medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Genomic architecture and prediction of censored time-to-event phenotypes with a Bayesian genome-wide analysis

Sven E. Ojavee[1,*], Athanasios Kousathanas[1], Daniel Trejo Banos[1], Etienne J. Orliac[2], Marion Patxot[1], Kristi Läll[3], Reedik Mägi[3], Krista Fischer[3,4], Zoltan Kutalik[5,6], Matthew R. Robinson[7,*]

1 Department of Computational Biology, University of Lausanne, Lausanne, Switzerland.
2 Scientific Computing and Research Support Unit, University of Lausanne, Lausanne, Switzerland.
3 Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
4 Institute of Mathematics and Statistics, University of Tartu, Tartu, Estonia
5 University Center for Primary Care and Public Health, Lausanne, Switzerland
6 Swiss Institute of Bioinformatics, Lausanne, Switzerland
7 Institute of Science and Technology Austria, Klosterneuburg, Austria.

*corresponding author: svenerik.ojavee@unil.ch or matthew.robinson@ist.ac.at

Here, we develop a Bayesian approach (BayesW) that provides probabilistic inference of the genetic architecture of age-at-diagnosis of disease and time-to-event phenotypes. We show in extensive simulation work that our method provides insight into genetic effects underlying disease progression, achieving higher statistical power and improved genomic prediction as compared to other approaches. We develop a hybrid-parallel sampling scheme facilitating age-at-onset analyses in large-scale biobank data. In the UK Biobank, we find evidence for an infinitesimal contribution of many thousands of common genomic regions to variation in the onset of common complex disorders of high blood pressure (HBP), cardiac disease (CAD), and type-2 diabetes (T2D), and for the genetic basis of age-at-onset reflecting the underlying genetic liability to disease. In contrast, while age-at-menopause and age-at-menarche are highly polygenic, we find higher variance contributed by low frequency variants. We find 360 independent 50kb regions for age-at-menarche with $\geq 95\%$ posterior inclusion probability of contributing 0.001% to the genetic variance, 115 regions for age-at-menopause, 246 regions for age-at-diagnosis of HBP, 32 regions for CAD, and 56 regions for T2D. Genomic prediction into the Estonian Genome Centre data shows that BayesW gives higher prediction accuracy than other approaches.

Genome-wide association studies (GWAS) have greatly expanded our understanding of the genetic architecture of complex-traits, but have largely focused on binary phenotypes and quantitative traits [1], leaving the age-at-onset of symptoms little studied, despite it being one of the key traits in biobank studies of age-related disease. Understanding the environmental and genetic basis of the time at which symptoms first occur is critical for early screening programs and for gaining insight into disease development and progression, especially as the pathological processes of many age-related diseases may be triggered decades before the first symptoms appear. Evidence suggests that genome-wide analyses conducted with case-control phenotypes tend to have less power in comparison with their age-at-onset analysis counterparts [2,3]. Genetic predictors created from case-control studies have been shown to be predictive of age-at-diagnosis [4], implying that early-onset is to a certain degree indicative of a higher underlying liability of disease. However, our understanding of the genetic architecture of reproductive timing, and the age at which symptoms first develop for common complex disorders, remains limited.

Statistical modelling of time-to-event data is a highly active research area, and is frequently applied to clinical and pharmacogenetic studies. Analogous to single marker regression in GWAS analyses, a Cox proportional hazards (PH) model [5] for each single nucleotide polymorphism (SNP) $j \in \{1, ..., M\}$ can be formulated as $h_i(t) = h_0(t)\exp(x_{ij}\beta_j)$, where $h_0(t)$ is the baseline hazard at time t, $h_i(t)$ is the hazard for individual i, x_{ij} is the standardised jth SNP marker value, with β_j the effect size of the jth SNP and M the total number of SNPs [6–8]. Recently there have been improvements in the computation times using some approximations for single-marker Cox PH regression [9], however this approach still yields marginal effect size estimates as the markers are not fitted simultaneously. Residual based approaches have also been widely used, which first regress the phenotype on covariates such as gender or age at entry in Cox

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

PH model, and then use the residuals in a second regression on the SNP data, with Martingale residuals $\hat{M}_i = d_i - \hat{\Lambda}_0(t_i)\exp(Z_i^t\gamma)$,where $\hat{M}_i$ is the residual for individual i, δ_i is the failure indicator ($d_i = 1$ for event during the study period, otherwise $d_i = 0$), $\hat{\Lambda}_0(t_i)$ is the baseline cumulative hazard function at time t_i, t_i is the follow-up time for individual i, Z_i is the vector of variables used in the first regression step and γ the vector of corresponding parameter estimates [10, 11]. The martingale residual approach retains the linearity between the effect and the phenotype and given the model in the second step it can also be very fast, however it combines the failure time and censoring indicator to one summary statistic which might diminish the power of this model. Rather than testing markers one-at-a-time, their effects can be estimated jointly in a mixed-effects Cox PH model, referred to as a frailty model, specified as $\lambda_i(t|b) = \lambda_0(t)\exp(X_i^t\beta + b)$, where β is the effect for one SNP being tested along with other fixed effects such as age or sex, $b \sim N(0, \sigma^2\Sigma)$ is the N-dimensional vector of random effects (N is the sample size), $\Sigma : N \times N$ is the genetic relationship matrix, σ^2 is the variance of the genetic component, $\lambda_0(t)$ is the baseline hazard function and $\lambda_i(t|b)$ is the hazard for individual i. This idea has been long limited by computational resources and in the latest implementation (COXMEG) [12] analyses are constrained to around $\sim 10,000$ individuals. For joint marker effect estimation, there is also the Cox LASSO model [13] which has been recently developed for genetic data in the R package snpnet [14, 15]. Fully parametric alternatives are also the Sparse Bayesian Weibull regression (SBWR), which may outperform LASSO based approaches [16], but like other Bayesian methods such as SurvEMVS [17] or a semi-parametric g-prior approach of Held et al [18] the ultrahigh dimensions of genetic data limit their application. Therefore, approaches that can efficiently handle both the complexity and scale of many millions of sequenced individuals with time-to-event outcomes have not been extensively developed, limiting our understanding and our ability to predict disease progression and the timing of symptom onset.

Here, we take an alternative approach to obtain accurate inference in full-scale phenotype-genotype sequence data sets, by proposing a mixture of regressions model with variable selection, using different regularisation parameters for genetically motivated groups (see Methods). Our suggested model fits all of the markers jointly in a Bayesian framework using Weibull assumption for the phenotypes. We show that this approach: (1) allows for a contrasting the genetic architectures of age-at-onset phenotypes under this flexible prior formulation; (2) yields marker effect estimates β_j that represent the effect of each marker conditional on the effects of all the other markers accounting for genetic architecture; (3) provides a determination of the probability that each marker and genomic region is associated with a phenotype, alongside the proportion of phenotypic variation contributed by each, and (4) gives a posterior predictive distribution for each individual. Regardless of the phenotypic distribution, our suggested approach greatly improves genomic prediction for the timing of events for each individual and enables better insight behind the genetic architecture underlying time-to-event traits.

BayesW model

An overview of our model is as follows, suppose that M markers are split between Φ different groups. The groups can be for example formed based on marker-specific genomic annotations, MAF grouping, grouping based on LD score etc. We assume for an individual i that the age-at-onset of a disease Y_i has Weibull distribution, with a reparametrisation of the model to represent the mean and the variance of the logarithm of the phenotype as

$$E(\log Y_i|\mu, \beta, \delta, \alpha) = \mu + \sum_{\varphi=1}^{\Phi} x_i^{\varphi}\beta^{\varphi} + z_i\delta, \tag{1}$$

$$Var(\log Y_i|\mu, \beta, \delta, \alpha) = \frac{\pi^2}{6\alpha^2}, \tag{2}$$

where μ is the intercept, x_i^{φ} are the standardised marker values for all SNPs in group φ, β^{φ} are the marker estimates for the corresponding group, z_i are additional covariate values (such as sex or genetic principal components), δ are the additional covariate effect estimates and α is the Weibull shape parameter (see Methods). For each group, we assume that β^{φ} are distributed according to a mixture of L_{φ} Gaussian components. Each marker (from group φ) estimate β_j $j \in \{1, ..., M\}$ is related to a corresponding indicator

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

variable $\gamma_j \in \{0, ..., L_\varphi\}$ where L_φ is the number of mixture distributions. β_j have zero values if and only if $\gamma_j = 0$. We assume that non-zero β_j, where marker j belongs to group φ, that has been assigned to mixture component ($\gamma_j = k \geq 1$) comes from a normal distribution with zero mean and variance $C_k^\varphi \sigma^2_{G\varphi}$, that is $\beta_j \sim N(0, C_k^\varphi \sigma^2_{G\varphi})$, where $\sigma^2_{G\varphi}$ represents the phenotypic variance attributable to markers of group φ and C_k^φ is a group and mixture specific factor showing the magnitude of variance explained by this specific mixture that is given by the user. For example, specifying $C_1^\varphi = 0.0001$, $C_2^\varphi = 0.001$ and $C_3^\varphi = 0.01$ gives us mixtures that respectively explain 0.01%, 0.1% and 1% of the genetic variance. We also assume that prior probabilities of belonging to each of the mixture distribution k is stored in $L_\varphi + 1$-dimensional vector π^φ. Thus the mixture proportions, variance explained by the SNP markers, and mixture constants are all unique and independent across SNP marker groups.

An algorithm for whole-genome data at biobank scale

We develop a computational framework that overcomes previous limitations for the application of age-at-onset models to large-scale biobank studies. In the model likelihood, we account for right censoring, a situation where only the last known time without an event is recorded, with the event potentially taking place some time in the future (see Methods). Although we did not apply it in our final analysis, we also formulate the model to accommodate left truncation, a situation where individuals are not missing from the data at random, creating differences in genetic composition of individuals across age groups (see Methods). We implement a parallel sampling scheme for Eq.(44) that allows the data to be split across compute nodes (in a series of MPI tasks), whilst still maintaining accuracy of the estimation of β_j. With T parallel workers (MPI tasks), Bulk Synchronous Parallel (BSP) Gibbs sampling can sample T marker effects when sequential Gibbs samples a single one, but BSP requires an extra synchronisation step of the tasks after each of them has processed u markers (see Methods). After each worker has processed u markers, we synchronise the workers by transmitting the residual vector across workers. Given our assumption that the phenotype follows a Weibull distribution, we are using a numerical method, Adaptive Gauss-Hermite quadrature, for calculating the mixture membership probabilities for variable selection, and Adaptive Rejection Sampling (ARS) for estimating the marker effects. We implement these approaches to take full advantage of the sparsity of genomic data, converting computationally intensive calculations of exponents and dot products into a series of summations. We provide publicly available software (see Code Availability) that has the capacity to easily extend to a wider range of models (not just Weibull) than that described here. Our software enables the estimation of 2,975,268 SNP inclusion probabilities split between $T = 96$ workers, using 12 compute nodes and synchronisation rate of $u = 10$, mixture allocation and effect sizes in 371,878 individuals with an average of 49.7 seconds per iteration without groups model and 50.0 seconds per iteration with the groups model; using 151,472 individuals with $T = 64$, 8 compute nodes and $u = 10$ without groups we get an average of 24.8 and with groups we get an average of 26.4 seconds per iteration. Here, we have chosen to run the chains for 10,000 iterations leading to execution times of 69 hours (N =151,472) to 139 hours (N =371,878). The run times ultimately depend upon compute cluster utilisation and the genetic architecture of the phenotype, as calls to the ARS procedure are linear with the number of markers. The calculations were done by using Helvetios cluster of EPFL (see Code availability).

Simulation study

We show in a simulation study that our model estimates SNP marker effect sizes more accurately, with higher power, and thus obtains improved genomic prediction accuracy as compared to other available methods (Figure 1, Figure S1 and S2). First, we show that this statement holds even in the case of model misspecification (Figure 1), where the phenotypic distribution does not correspond to a Weibull distribution, but rather conforms to a series of different generalized gamma distributions (of which Weibull is one), with differing theta value (see Methods, and Supplementary Note). In a simulation study of $N = 5,000$ individuals and 50,000 uncorrelated markers with $p = 500$ randomly selected markers as causal variants, our BayesW model obtains higher out of sample prediction accuracy than a Cox LASSO, or a Martingale residual approach used in a number of recent studies (Figure 1a). Second, we show that this statement also holds in a larger simulation study using a real genomic data-set of $N = 20,000$ randomly selected UK Biobank individuals

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

and 194,922 correlated genetic markers on chromosome 22 under different censoring levels (Figure S1). Our approach achieves better precision-recall as compared to these approaches (Figure 1b) across all values of theta and all censoring levels within the data (Figure S2). We show that across the range of theta values that generate model misspecification, SNP marker effect estimates remain correctly estimated, but that a background of small-effect false positives enter the model, creating an underestimate of the true effect sizes (Figure 1c). Third, in the Supplementary Note we derive a definition of SNP heritability, the proportion of among-individual variation in age-at-onset that is attributable to SNP effects, for both the variance of the logarithmed phenotype and the original scale (see Supplementary Note). We show that the log-scale SNP heritability definition is valid under a Weibull assumption and across the range of theta when restricting the markers entering the model (Figure 1d, single mixtures 0.01), but may be inflated under low theta values (Figure 1d, mixtures 0.001, 0.01) because of the increase in small-effect false positives that enter the model (Figure 1e). We show below that these false positives are controlled for under all generative models when conducting variable selection using a posterior inclusion probability threshold of $\geq 95\%$. Finally, we show that our BSG algorithm is stable under a wide range of synchronisation rates, parallelism, and quadrature point selection (Figure S3).

The genetic architecture of age-at-onset

We then applied our model to unrelated UK biobank individuals of European ancestry with a pruned set of M=2,975,268 SNPs for five traits: two reproductive phenotypes of age-at-menopause (N =151,472) and age-at-menarche (N =200,493) and three common complex diseases (selected as they are some of leading causes of mortality) of time-to-diagnosis of type-2-diabetes (T2D) (N =372,280), coronary-artery-disease (CAD) (N =360,715) and high blood pressure (HBP) (N =371,878). Using our BSG sampling scheme, we ran a baseline model without any grouping of markers, and then we re-ran the model grouping markers into 20 MAF-LD bins (quintiles of MAF and then quartiles within each MAF quantile split by LD score). Groups were defined using MAF and LD based on recent theory [19] and recent simulation study results [20–23], which suggest that accurate estimation of genetic variance might require accounting for the MAF-LD structure. In order to understand the effect size distribution and genetic architecture, four mixture components were specified such that they would represent 0.001%, 0.01%, 0.1% or 1% of the total genetic variance for the no groups model and for the groups model the mixtures would represent 0.01%, 0.1%,1% or 10% of the group-specific genetic variance. Additional variables such as sex, UK Biobank assessment centre, genotype chip, and the leading 20 PCs of the SNP data (see Methods) were used as fixed effects in the analysis. We conducted a series of convergence diagnostic analyses of the posterior distributions to ensure we obtained estimates from a converged set of chains (Figures S4, S5, S6 and S7).

Under the assumption that the traits are Weibull distributed, this gives log-scale SNP (pseudo-)heritability estimates (see Supplementary Note) of 0.26 (95% CI 0.25, 0.27) for age-at-menopause, 0.41 (95% CI 0.40, 0.42) for age-at-menarche, 0.36 (95% CI 0.35, 0.37) for age-at-diagnosis of HBP, 0.48 (95% CI 0.44, 0.52) for age-at-diagnosis of CAD, and 0.52 (95% CI 0.50, 0.55) for age-at-diagnosis of T2D. Both the model with and without groups reach similar conclusions in terms of partitioning markers between mixtures (Figure 2a) indicating that the inference we draw on the genetic architecture is here independent of the group-specific prior specification. However, our BayesW grouped mixture of regression model allows for contrasting the variance contributed by different MAF and LD groups across traits. For all traits, we find that the majority of the variance contributed by SNP markers is attributable to SNPs that each proportionally contribute an average of 10^{-5} of the genetic variance (Figure 2a). We find evidence that age-at-menarche is highly polygenic with 88.1% (95% CI 86.8%, 89.4%) of the genetic variance attributable to the SNPs contributed by markers in the 10^{-5} mixture group, similar to CAD with 74.2% (95% CI 63.6%, 81.5%, Figure 2b). Age-at-menopause and age-at-T2D diagnosis are the least polygenic, with 32.3% (95% CI 28.9%, 35.7%) and 18.9% (95% CI 14.6%, 22.9%) of the genotypic variance attributable to the SNPs contributed by markers in the 10^{-3} mixture group respectively (Figure 2b).

We find marked differences in the underlying genetic architecture of these different age-at-onset phenotypes (Figure 2c,d). For age-at-menarche, many rare low-LD SNPs and many common SNPs contribute similar proportions to the phenotypic variance attributable to the SNP markers, implying larger absolute effect sizes for rare low-LD variants per minor allele substitution, with age-at-menopause showing a similar but

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

less pronounced pattern with a large proportion of the genetic variance stemming from moderate effect size common variants (Figure 2c). In contrast, we find evidence that the phenotypic variance attributable to the SNP markers for age-at-diagnosis for CAD, HBP, and T2D is predominantly contributed by common variants of small effect (Figure 2c). This implies that female reproductive traits may have been under far stronger selection in our evolutionary past than age-at-diagnosis of modern day common complex disease [24]. In summary, we find that most of the phenotypic variance attributable to SNPs is contributed by very many small effect common variants, but that there are key differences among time-to-event phenotypes, with reproductive traits showing different patterns of genetic architecture to time-to-diagnosis phenotypes.

We then partitioned the variance attributed to SNP markers across 50kb regions of the genome and determined the posterior inclusion probability that each region contributes at least 1/1000, 1/10,000, or 1/100,000 of the genotypic variance, providing a probabilistic approach to assess the contribution of different genomic regions to time-to-event phenotypes. We denote this as the posterior probability of the window variance (PPWV) [25] (see Methods) and we find 360 50kb regions for age-at-menarche with $\geq 95\%$ PPWV of 1/100,000, 115 regions for age-at-menopause, 246 regions for age-at-diagnosis of HBP, 32 regions for age-at-diagnosis of CAD, and 56 regions for age-at-diagnosis of T2D from our BayesW grouped mixture of regression model (Figure 3a). Our grouped model provides slightly higher power, as reflected by higher posterior inclusion probabilities at smaller effect sizes (Figures 3a, S8), with the baseline BayesW mixture of regression model detecting 31.0% fewer 50kb regions for age-at-menarche, 8.1% fewer for age-at-menopause, 2.6% fewer for age-at-diagnosis of HBP, 7.7% fewer for age-at-diagnosis of CAD, and 18.7% fewer for age-at-diagnosis of T2D when using 1/10,000 PPWV threshold. Similarly we evaluated region-based significance by calculating PPWV for regions that were created by mapping markers to the closest gene (Figure 3b). For age-at-menopause we find 101, for age-at-menarche we find 119, for time-to-T2D we find 41, for time-to-HBP we find 159 and for time-to-CAD we find 20 gene-associated regions with $\geq 95\%$ PPWV of explaining at least 1/10,000 of the genetic variance. In addition, we find evidence for differences in the effect size distribution across traits, largely reflecting differences in power that result from sample size differences and different censoring levels across traits (Figure 3c,d) (see Methods). Overall, these results suggest many hundreds of genomic regions spread throughout the genome contribute to the timing of common complex traits.

We also compared the 50kb regions discovered by BayesW with 50kb regions discovered by another association method fastGWA [26]. Although for age-at-menarche and age-at-menopause we find 109 and 62 50kb regions that are concordantly significant according to the two methods (Table 1), we find less concordance among the other traits. For time-to-angina and -heart attack fastGWA does not find any significant regions, for time-to-HBP BayesW finds greatly more 50kb regions (BayesW: 246, fastGWA: 35). The striking difference between the numbers of identified 50kb regions could be largely attributed to the larger sample size of BayesW as BayesW can also use the data from censored individuals where fastGWA can only resort to the uncensored individuals. For time-to-diabetes the number of identified regions is similar but almost all of the 50kb regions are different. Although for time-to-menarche and time-to-menopause the fastGWA identifies more 50kb regions, still around half (menopause) or two thirds (menarche) of the regions identified by BayesW are not picked up by fastGWA. Moreover, the BayesW shows greatly better prediction accuracy to Estonian biobank compared to fastGWA when predicting age-at-menarche or age-at-menopause (Figure 4a,b) indicating that the regions identified by BayesW are reflecting the genetic architecture more accurately. Therefore, BayesW identifies novel regions compared to previous association methods; for time-to-diagnosis traits, it can discover more regions due to using the censored individuals; and the regions discovered by BayesW result in greatly improved prediction accuracy compared to fastGWA.

Out of sample prediction in an Estonian population

We used the estimates obtained from the group-specific model to predict time-to-event in $N = 32,594$ individuals of the Estonian Biobank data. We compared our model performance to the snpnet Cox-LASSO approach [14, 15] trained on the same UK Biobank data (see Methods) using two metrics. As some of the Estonian Biobank time-to-event phenotypes are censored, we choose to calculate the R^2 values between the predicted values and the martingale residuals from the Cox PH model where the true phenotypes are regressed on sex. In addition we calculate Harrell's C-statistic [27] from the Cox PH model where the true phenotypic values are regressed on the predicted values and sex. BayesW outperforms snpnet for all

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Phenotype	BayesW: no fastGWA: no	BayesW: no fastGWA: yes	BayesW: yes fastGWA: no	BayesW: yes fastGWA: yes
Time to Angina	52797	0	32	0
Time to Heart attack	52797	0	32	0
Time to HBP	52558	25	236	10
Time to Menarche	51545	924	251	109
Time to Menopause	52292	422	53	62
Time to Diabetes	52702	70	55	1

Table 1. Concordance between the 50kb regions discovered by BayesW or fastGWA. We split the genome into 50kb regions and we evaluated the significance of each of the regions with using the results from the groups BayesW model and the fastGWA model. The fastGWA results for our CAD and T2D definition were missing so instead time-to-angina and time-to-heart attack are shown for CAD and time-to-diabetes is shown for T2D. Here, BayesW calls a 50kb region significant if the PPWV of the region (explaining at least 0.001% of the genetic variance) is higher than 0.95; fastGWA calls a 50kb region significant if there exists at least one marker with a p-value $< 5 \cdot 10^{-8}$. We find that although for age-at-menarche and age-at-menopause there exists an abundance of 50kb regions with concordant significance, for other traits most of the discovered 50kb regions differ between two methods.

phenotypes (Figure 4 a,b) by giving R^2 of 0.032 compared to snpnet's 0.017 for age-at-menopause of 18,134 women and 0.05 compared to snpnet's 0.040 for age-at-menarche of 18,134 women. We also get an increase in Harrell's C-statistic with BayesW giving 0.623 (se = 0.00443) compared to snpnet's 0.593 (0.00455) for age-at-menopause and for age-at-menarche we get C-statistic of 0.598 (0.00290) with BayesW compared to snpnet's 0.580 (0.00294). For the age-at-diagnosis traits, we obtain R^2 values of 0.0047, 0.0236, and 0.0441 for BayesW and 0.0030, 0.0135, and 0.0271 with snpnet for CAD, T2D, and HBP respectively. This shows that our BayesW model gives a higher prediction accuracy compared to the snpnet method, in-line with our simulation study results.

We then compared the BayesW prediction results to those obtained from a case-control analysis. In a companion paper [19], we develop the BayesRR-RC approach and we use this to analyse the indicator variable (0 = no registered disease, 1 = reported disease) for HBP using the same data and a liability model to facilitate a direct comparison of the methods. For CAD and T2D, we use the results of the companion paper, where there were almost twice as many case observations (for CAD BayesRR-RC had 39,776 vs BayesW 17,452 and for T2D BayesRR-RC had 25,773 vs BayesW 15,813 cases) as it included those with confirmed diagnosis but no age information and 8.4 million SNPs were analysed. For the prediction of age-at-diagnosis, we compared the R^2 values between the predicted values and the martingale residuals from the Cox PH model and the Harrell's C-statistic (Figure 4a,b). For HBP, CAD we find that BayesW marginally outperforms BayesRR-RC with (HBP R^2 BayesW 0.0441, BayesR 0.0437; CAD R^2 BayesW 0.0047, BayesR 0.0046) and for T2D BayesRR-RC marginally outperforms BayesW (R^2 BayesW 0.0236, BayesR 0.0262). Similar ranking can be observed when using Harrell's C-statistic for comparison (Figure 4b). We then compared approaches when predicting 0/1 case-control status, rather than age-at-diagnosis (Figure 4c,d). We find that for predicting HBP BayesW marginally outperforms BayesRR-RC in terms of R^2 (BayesW 0.0375, BayesR 0.0365) and area under PR curve (BayesW 0.339, BayesR 0.336) (used because of the imbalance between cases and controls); for predicting CAD or T2D despite the increase case sample size, BayesRR-RC only marginally outperforms BayesW (T2D R^2 BayesW 0.0127, BayesR 0.0136 and AUC BayesW 0.0766, BayesR 0.0799; CAD R^2 BayesW 0.0025, BayesR 0.0027 and AUC BayesW 0.0920, BayesR 0.0941). Therefore we get very similar prediction accuracies with both methods when predicting case-control phenotypes although the BayesW model was estimated using time-to-event phenotypes with less cases for T2D and CAD.

A finding of similar prediction accuracy is unsurprising given the striking concordance between the results of the two models when partitioning the genotype into 50kb regions. We calculated (on logarithmic scale) the mean proportion of genetic variance attributed to each 50kb region for the model using case-control phenotype (BayesRR) and for the model using time-to-event phenotype (BayesW). Both models attribute similar amounts of genetic variance to the same 50kb regions (Figure S9), with correlations between logarithmic proportions of genetic variances are 0.941, 0.647 and 0.554 for HBP, T2D and CAD respectively. Thus, we have shown here that using either time-to-event or case-control data for genome-wide association analysis we find similar amount of genetic variance attributed to the same regions and both analyses have similar

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

predictive performance when predicting case-control phenotypes. This suggests that to some extent there is interchangeability between the case-control and time-to-event phenotypes demonstrating that both phenotypes are describing a similar latent mechanism.

The BayesW model enables posterior predictive distributions to be generated for each individual. For evaluating the predictive performance of BayesW on the reproductive traits, we calculated the 95% credibility prediction intervals for each of the subjects in the Estonian biobank. We chose to evaluate reproductive traits only as it is sure that every woman should experience those events given they have reached a sufficient age. For age-at-menopause 92.3% and for age-at-menarche 94.8% of the true uncensored phenotypes lie within the BayesW 95% credibility prediction intervals. This demonstrates that even though the prediction R^2 values for those traits are not extremely high due to the low genetic variance underlying the phenotypic variance, our Bayesian model quantifies the model uncertainty and yields well-calibrated subject-specific prediction intervals. An example of the shape of the distribution is shown in the Figure S11. A caveat is that the subject-specific prediction intervals can be rather wide, though approximately half the width of the data range. For age-at-menopause the data range from 34 to 63 (a width of 29 years), and the width of the 95% credibility intervals ranged from 13.4 to 18.6 years with a median width of 15.9 years. For age-at-menarche, data values range from 9 to 19, and the posterior predictive interval width ranged from 5.1 to 7.9 years with a median width of 6.3 years.

Discussion

Here, we have shown that our BayesW mixture of regressions model provides inference as to the genetic architecture of reproductive timing and the age at which symptoms first develop for common complex disorders. We provide evidence for an infinitesimal contribution of many thousands of common genomic regions to variation in the onset of common complex disorders and for the genetic basis of age-at-onset reflecting the underlying genetic liability to disease. In contrast, while age-at-menopause and age-at-menarche are highly polygenic, average effects sizes and the variance contributed by low frequency variants is higher, supporting a history of stronger selection pressure in the human population [24].

Genome-wide association studies of time-to-event phenotypes are critical for gaining insights into the genetics of disease etiology and progression, but their application has been hampered by the computational and statistical challenges of the models, especially when the predictors are ultrahigh-dimensional. Our hybrid-parallel sampling scheme enables a fully Bayesian model to be applied to large-scale genomic data and we show improved genomic prediction over competing approaches, not only in the R^2 or C-statistic obtained, but in the inference that can be obtained from a full posterior predictive distribution. Previous evidence shows that cohort studies using proportional hazards (or Cox) regression models generally increase statistical power compared to case-control studies using logistic regression model [2, 3]. Our results support this and we expect the benefits to become more evident as the number of cases accrue with accurate age-at-diagnosis information.

A typical approach in time-to-event analysis is the Cox PH model [5] that uses a non-parametric estimate for the baseline hazard and then estimates other effect sizes proportional to this hazard. Our BayesW model is also a proportional hazard model with the constraint that the baseline hazard follows a Weibull distribution and thus marker effect size estimates have similar interpretation as those from a Cox PH model. Interestingly, the results from both our simulation study and real data analysis show that when quantifying the significance of the markers and estimating the marker effect sizes, it might not be pivotal to capture the baseline hazard with a non-parametric method. The simulations show that even in the misspecified cases BayesW performs better compared to the semi-parametric Cox model, demonstrating that using a parametric assumption might be more descriptive than simply using a Cox PH model from standard practice.

There has been a significant amount of work on the heritability of the time-to-event traits. For example, it has been suggested to define heritability in the Weibull frailty model on the log-time scale [28], or on the log-frailty scale in Cox PH model [29]. Transforming the log-scale heritability to the original scale [30], has then required approximations and the term of original scale heritability has not been easy to explain and use [31]. Here, using a similar idea of partitioning the total phenotypic variance into genetic and error variance components, we present an expression for SNP heritability on the log-time scale. We then show that there exists a natural correspondence between log-scale and original scale heritabilities, without the need

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

for any approximations, with log-scale and original scale heritability giving similar estimates if the Weibull shape parameter tends to higher values. Therefore, under Weibull assumptions, we provide a definition of SNP-heritability for time-to-event traits for the GWAS-era.

There are a number of key considerations and limitations. The assumption of a Weibull distribution for the traits considered here can induce bias in the hyperparameter estimates, although we have shown that this assumption yields accurate results in terms of prediction regardless of the phenotypic distribution. A third parameter could be introduced through the use of a generalised gamma distribution and this will be the focus of future work as it should allow for unbiased hyperparameter estimation irrespective of the trait distribution. Additionally, despite allowing for left-truncation in the likelihood, we focus on presenting a series of baseline results before extending our inference to account for differences in sampling, semi-competing risks across different outcomes, genomic annotation enrichment, and sex-differences both the effect sizes and in the sampling of different time-to-event outcomes all of which require extensions to the modelling framework, which are also the focus of future work. Furthermore, we do not consider time-varying coefficients or time-varying covariates, which may improve inference as multiple measurements over time are collected in biobank studies. Nevertheless, this work represents a first step toward large-scale inference of the genomic basis of variation in the timing of common complex traits.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

Parametrisation of Weibull distribution

We define Y_i as the time-to-event for an individual i, with Weibull distribution $Y_i \sim W(a, b_i)$, where a and b_i are correspondingly the shape and scale parameters. The survival function is

$$S_i(y) = \exp\Big(-\Big(\frac{y}{b_i}\Big)^a\Big). \tag{3}$$

We are interested in modelling the mean and the variance of the time-to-event. Unfortunately, the mean and the variance of Weibull are dependent as they share both parameters in their expressions. Moreover, as the expressions for mean and variance contain gamma functions it is rather difficult to dissect the mean and variance to be dependant only on one parameter. One possible solution is to use $\log Y_i$ and its moments instead. If $Y_i \sim W(a, b_i)$ then $\log Y_i$ has a Gumbel distribution, with mean and variance

$$E(\log Y_i|b_i, a) = \log b_i - \frac{K}{a}, \tag{4}$$

$$Var(\log Y_i|b_i, a) = \frac{\pi^2}{6a^2}. \tag{5}$$

where K is the Euler-Mascheroni constant, $K \approx 0.57721$. The parametrisation for the variance is only dependent on one parameter which we denote as $\alpha = a$. As we are interested in modelling SNP effects β, covariates δ (sex, PCs) and the average scale for time-to-event μ (intercept), it is possible to introduce them in the following way $b_i = \exp\Big(\mu + x_i\beta + z_i\delta + \frac{K}{\alpha}\Big)$, resulting in

$$E(\log Y_i|\mu, \beta, \delta, \alpha) = \mu + x_i\beta + z_i\delta, \tag{6}$$

$$Var(\log Y_i|\mu, \beta, \delta, \alpha) = \frac{\pi^2}{6\alpha^2}, \tag{7}$$

where x_i is the transposed vector of scaled SNP marker values and z_i is the transposed vector of covariate values for an individual i and $\pi = 3.14159...$ is a constant. The third and the fourth moment for $\log Y_i$ are constant regardless of the parametrisation.

Modelling time-to-event and age-at-onset

As a baseline model, we propose to test association of Y_i with a series of covariates (SNP markers in this case) X using a mixture of regression model, with γ_j as the mixture indicator, with $\gamma_j = k$ if jth marker is included into the kth mixture component of the model, $k \in \{1, ..., L\}$, and $\gamma_j = 0$ if it is not included into the model. The expected value of time-to-event logarithm is then a linear combination of the markers included into the model plus the effect of the covariates and the intercept (μ).

$$E(\log Y_i|\mu, \beta, \delta, \alpha) = \mu + x_i\beta + z_i\delta, \tag{8}$$

$$Var(\log Y_i|\mu, \beta, \delta, \alpha) = \frac{\pi^2}{6\alpha^2}, \tag{9}$$

β_j have non-zero values if and only if $\gamma_j \geq 1$. We assume that non-zero β_j from mixture component $k > 0$ ($\gamma_j = k$) come from a normal distribution with zero mean and variance $C_k\sigma_G^2$, that is $\beta_j \sim N(0, C_k\sigma_G^2)$.

The survival and density function for Y_i are correspondingly

$$S_i(y) = \exp\{-y^\alpha \exp(-\alpha(\mu + x_i\beta + z_i\delta) - K)\}, \tag{10}$$

$$f_i(y) = \exp\{K - \alpha(\mu + x_i\beta + z_i\delta) - y^\alpha \exp(-\alpha(\mu + x_i\beta + z_i\delta) - K)\}y^{\alpha-1}\alpha, \tag{11}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

The likelihood function for the right censored and left truncated data of n individuals is then

$$p(D|\alpha,\beta,\delta,\mu) = \prod_{i=1}^{n} \frac{1}{S(a_i)} \prod_{i=1}^{n} f(y_i)^{d_i} S(y_i)^{1-d_i} =$$
$$\alpha^d \exp\{-Kd + (\alpha-1)\sum_{i=1}^{n} d_i \log y_i - \alpha \sum_{i=1}^{n} d_i(\mu + x_i\beta + z_i\delta) -$$
$$- \sum_{i=1}^{n} y_i^{\alpha} \exp(-\alpha(\mu + x_i\beta + z_i\delta) - K) + \sum_{i=1}^{n} a_i^{\alpha} \exp(-\alpha(\mu + x_i\beta + z_i\delta) - K)\}, \quad (12)$$

where d_i is the failure indicator and d is the number of events in the end of the periods; a_i is the time of left truncation.It is possible to use the model without left truncation. In order to do so, for every i, we will assume that $a_i = 0$. Whenever $a_i = 0$, we will naturally define $\exp(\alpha(\log(a_i) - \mu - x_i\beta - z_i\delta)) = 0$, thus the left truncation would not contribute to the likelihood in the equation 12.

Let the prior distribution of α be a gamma distribution with parameters α_0, κ_0

$$p(\alpha) \propto \alpha^{\alpha_0 - 1} \exp(-\kappa_0 \alpha), \quad (13)$$

the prior for β_j be normal:

$$p(\beta_j|\sigma_G^2, \gamma_j = k) \propto (\frac{1}{C_k \sigma_G^2})^{0.5} \exp[-\frac{1}{2C_k\sigma_G^2}\beta_j^2], \quad (14)$$

the prior for σ_G^2 be inverse gamma distribution with parameters α_σ and β_σ

$$p(\sigma_G^2) \propto (\frac{1}{\sigma_G^2})^{\alpha_\sigma + 1} \exp(-\frac{\beta_\sigma}{\sigma_G^2}), \quad (15)$$

the prior for δ_q (qth covariate) be normal with variance parameter σ_δ^2:

$$p(\delta_q) \propto \exp(-\frac{1}{2\sigma_\delta^2}\delta_q^2), \quad (16)$$

the prior for μ be normal with variance parameter σ_μ^2:

$$p(\mu) \propto \exp(-\frac{1}{2\sigma_\mu^2}\mu^2), \quad (17)$$

the prior for γ_j be multinomial:

$$p(\gamma_j|\pi) = \pi_0^{I(\gamma_j=0)} \cdot \ldots \cdot \pi_L^{I(\gamma_j=L)}, \quad (18)$$

the prior probabilities of belonging to each of the mixture distributions k are stored in $L+1$-dimensional vector π with the prior for π a Dirichlet distribution

$$p(\pi) = \text{Dirichlet}(\mathbf{p}_\text{L}), \quad (19)$$

where $I(\cdot)$ is the indicator function and $\mathbf{p}_\text{L}$ is the $L+1$-dimensional vector with prior values.

The conditional posterior distribution for σ_G^2 is inverse gamma with parameters

$$\alpha_\sigma + 0.5\sum_{k=1}^{L} |\gamma^k| \text{ and } \sum_{k=1}^{L} \beta'_{\gamma^k}\beta_{\gamma^k} + \beta_\sigma,$$

where γ^k denotes the set of indices j for which $\gamma_j = k$. The conditional posterior distribution for π is Dirichlet distribution

$$p(\pi|\gamma) = \text{Dirichlet}(\mathbf{p}_\text{L} + (|\gamma^0|, \ldots, |\gamma^L|)). \quad (20)$$

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Unfortunately, there is no simple form for the conditional posteriors of α, μ, β_j and δ_q. However, the conditional posterior distributions are log-concave (see Supplementary Note), and thus the sampling for α, μ, β_j and δ_q can be conducted using adaptive rejection sampling requiring only the log posteriors. Denoting β_{-j} as all the β parameters, excluding β_j, and δ_{-q} as all the δ parameters, excluding δ_q, these are

$$\log p(\alpha|D,\mu,\beta,\delta) = const + (\alpha_0 + d - 1)\log\alpha + \alpha[\sum_{i=1}^{n} d_i(\log y_i - \mu - x_i\beta - z_i\delta) - \kappa_0] + \exp(-K)\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta - z_i\delta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta - z_i\delta))], \quad (21)$$

$$\log p(\beta_j|D,\alpha,\mu,\beta_{-j},\delta,\sigma_G^2,\gamma_j = k) = \log\left(\frac{const}{\sqrt{C_k}}\right) - \alpha\beta_j\sum_{i=1}^{n} d_i x_{ij} + \exp(-K)\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta - z_i\delta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta - z_i\delta))] - \frac{1}{2C_k\sigma_G^2}\beta_j^2, \quad (22)$$

$$\log p(\delta_q|D,\alpha,\mu,\beta,\delta_{-q},\sigma_G^2) = -\alpha\delta_q\sum_{i=1}^{n} d_i z_{ij} + \exp(-K)\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta - z_i\delta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta - z_i\delta))] - \frac{1}{2\sigma_\delta^2}\delta_q^2, \quad (23)$$

$$\log p(\mu|D,\alpha,\beta,\delta) = const - \alpha\mu d + \exp(-K)\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta - z_i\delta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta - z_i\delta))] - \frac{1}{2\sigma_\mu^2}\mu^2. \quad (24)$$

Selection of the mixture component

We intend to do variable selection and select mixture component by using the idea of spike and slab priors [32], where the spike part of the prior has a point mass on 0. SNP will be assigned to a mixture component by comparing the ratios of marginal likelihood. For mixture selection for the jth SNP, we need to find the following marginal likelihood for every k. Suppose here that $C_0 > 0$ is the factor for 0th mixture (spike)

$$p(D|\beta_{-j},\delta,\sigma_G^2,\alpha,\mu,\gamma_j = k) = \int_{\beta_j} p(D|\beta_{-j},\delta,\mu,\alpha,\beta_j)p(\beta_j|\sigma_G^2,\gamma_j = k)d\beta_j = \frac{Q}{\sqrt{C_k}}\int_{\beta_j}\exp\{h_k(\beta_j)\}d\beta_j, \quad (25)$$

where D represents the observed data, Q is a positive constant that is not dependent on k and

$$h_k(\beta_j) = -\alpha\beta_j\sum_{i=1}^{n} d_i x_{ij} + \exp(-K)\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta - z_i\delta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta - z_i\delta))] - \frac{1}{2C_k\sigma_G^2}\beta_j^2. \quad (26)$$

The probability for γ_j is :

$$p(\gamma_j = k|D,\beta_{-j},\delta,\sigma_G^2,\alpha,\mu) = Cp(D|\beta_{-j},\delta,\gamma_j = k,\sigma_G^2,\alpha,\mu)p(\gamma_j = k). \quad (27)$$

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

where C is a positive constant that is not dependent on k. Denoting $\Theta = \{D, \beta_{-j}, \delta, \sigma_G^2, \alpha, \mu\}$, the probability to include SNP j in the component k can be calculated as

$$p(\gamma_j = k|\Theta) = \frac{Cp(\gamma_j = k|\Theta)}{C(p(\gamma_j = 0|\Theta) + ... + p(\gamma_j = L|\Theta))}. \tag{28}$$

For every k

$$p(\gamma_j = k|\Theta) = \frac{\frac{\pi_k}{\sqrt{C_k}} \int_{\beta_j} \exp(h_k(\beta_j))d\beta_j}{\frac{\pi_0}{\sqrt{C_0}} \int_{\beta_j} \exp(h_0(\beta_j))d\beta_j + ... + \frac{\pi_L}{\sqrt{C_L}} \int_{\beta_j} \exp(h_L(\beta_j))d\beta_j} \tag{29}$$

Here, the numerator represents the marginal likelihood assuming jth variable is included to the kth mixture component.

In general it is not possible to find an analytic expression for the integrals presented in equation 25, thus some numeric method has to be used for approximating their values. For this we use adaptive Gauss-Hermite quadrature as the integral is improper with infinite endpoints.

We start by rewriting the expression 26 as

$$h_k(\beta_j) = -\alpha\beta_j \sum_{i=1}^{n} d_i x_{ij} + \sum_{i=1}^{n} [\exp(u_i - \alpha x_{ij}\beta_j) - \exp(v_i - \alpha x_{ij}\beta_j)] - \frac{1}{2C_k\sigma_G^2}\beta_j^2, \tag{30}$$

where $v_i = \alpha(\log y_i - \mu - x_{i,-j}\beta_{-j} - z_i\delta) - K$ and u_i is analogous with a_i instead of y_i. We introduce a reparameterisation with variable s

$$s = \frac{\beta_j}{\sqrt{2C_k\sigma_G^2}} \tag{31}$$

and therefore we get from equation 25

$$\begin{aligned}
&Q \int \frac{1}{\sqrt{C_k}} \exp\{h_k(\beta_j)\}d\beta_j = \\
&Q \int \frac{\sqrt{2C_k\sigma_G^2}}{\sqrt{C_k}} \exp\{-\alpha s\sqrt{2C_k\sigma_G^2} \sum_{i=1}^{n} d_i x_{ij} + \sum_{i=1}^{n} [\exp(u_i - \alpha x_{ij} s\sqrt{2C_k\sigma_G^2}) - \exp(v_i - \alpha x_{ij} s\sqrt{2C_k\sigma_G^2})] - s^2\}ds = \\
&Q\sqrt{2\sigma_G^2} \int \exp\{-\alpha s\sqrt{2C_k\sigma_G^2} \sum_{i=1}^{n} d_i x_{ij} + \sum_{i=1}^{n} [\exp(u_i - \alpha x_{ij} s\sqrt{2C_k\sigma_G^2}) - \exp(v_i - \alpha x_{ij} s\sqrt{2C_k\sigma_G^2})] - s^2\}ds = \\
&Q\sqrt{2\sigma_G^2} \int g_k(s)ds.
\end{aligned} \tag{32}$$

in the last expression in equation 29, the term $Q\sqrt{2\sigma_G^2}$ cancels out from the numerator and the denominator.

If the smallest mixture variance factor $C_0 > 0$, then the corresponding spike distribution is absolutely continuous. As we would like to use Dirac spike instead, we define the corresponding marginal likelihood by finding the limit of the expression in the process $C_0 \to 0+$.

$$p(D|\beta_{-j}, \delta, \sigma_G^2, \alpha, \mu, \gamma_j = 0) = \lim_{C_0 \to 0+} Q \int \frac{1}{\sqrt{C_0}} \exp\{h_0(\beta_j)\}d\beta_j. \tag{33}$$

We are only interested in C_0 in the limiting process so without the loss of generality we define C_0 through an auxiliary positive integer variable l as $C_0 = \frac{1}{l}$ and using the reparametrisation result from equation 32 we get that

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

$$p(D|\beta_{-j},\delta,\sigma_G^2,\alpha,\mu,\gamma_j=0)=$$
$$\lim_{C_0\to 0+} Q\sqrt{2\sigma_G^2}\int \exp\{-\alpha s\sqrt{2C_0\sigma_G^2}\sum_{i=1}^n d_i x_{ij}+\sum_{i=1}^n[\exp(u_i-\alpha x_{ij}s\sqrt{2C_0\sigma_G^2})-\exp(v_i-\alpha x_{ij}s\sqrt{2C_0\sigma_G^2})]-s^2\}ds=$$
$$\lim_{l\to\infty} Q\sqrt{2\sigma_G^2}\int \exp\{-\alpha s\sqrt{2\sigma_G^2/l}\sum_{i=1}^n d_i x_{ij}+\sum_{i=1}^n[\exp(u_i-\alpha x_{ij}s\sqrt{2\sigma_G^2/l})-\exp(v_i-\alpha x_{ij}s\sqrt{2\sigma_G^2/l})]-s^2\}ds=$$
$$\lim_{l\to\infty} Q\sqrt{2\sigma_G^2}\int f(s,l)\exp\{-s^2\}ds. \quad (34)$$

As $f(s,l)\leq 1$ for every possible combination of arguments, because in the data censoring or event occurs only after entering the study, we can write that

$$f(s,l)\exp\{-s^2\}\leq \exp\{-s^2\},\forall l \quad (35)$$

which means that the integrand in equation 34 is dominated by $\exp\{-s^2\}$. Furthermore, we see that the limit of the integrand is

$$\lim_{l\to\infty} f(s,l)\exp\{-s^2\}=\exp\{\sum_{i=1}^n[\exp(u_i)-\exp(v_i)]-s^2\}. \quad (36)$$

As $\int \exp\{-s^2\}ds<\infty$, it is possible to use the Lebesgue's dominated convergence theorem and therefore

$$\lim_{l\to\infty} Q\sqrt{2\sigma_G^2}\int f(s,l)\exp\{-s^2\}ds = Q\sqrt{2\sigma_G^2}\int \exp\{\sum_{i=1}^n[\exp(u_i)-\exp(v_i)]-s^2\}ds=$$
$$Q\sqrt{2\sigma_G^2}\exp\{\sum_{i=1}^n[\exp(u_i)-\exp(v_i)]\}\int \exp\{-s^2\}ds = Q\sqrt{2\pi\sigma_G^2}\exp\{\sum_{i=1}^n[\exp(u_i)-\exp(v_i)]\}. \quad (37)$$

In conclusion, we have derived the expression for the marginal likelihood for the Dirac spike variance component as

$$p(D|\beta_{-j},\delta,\sigma_G^2,\alpha,\mu,\gamma_j=0)=Q\sqrt{2\pi\sigma_G^2}\exp\{\sum_{i=1}^n[\exp(u_i)-\exp(v_i)]\}. \quad (38)$$

Adaptive Gauss-Hermite quadrature

It is possible to use Gauss-Hermite quadrature, however it can happen that for adequate precision one has to use large number of quadrature points leading to more calculations. Adaptive Gauss-Hermite quadrature can make the procedure more efficient. For any function $g_k(s)$ as defined in equation 32 we can write

$$\int_{-\infty}^{\infty} g_k(s)ds \approx \hat{\sigma}\sqrt{2}\sum_{r=1}^m w_r g_k(\hat{\mu}+\hat{\sigma}\sqrt{2}t_r), \quad (39)$$

where $\hat{\mu}$ could be chosen as the mode of $g_k(s)$ and $\hat{\sigma}=\frac{1}{\sqrt{-(\log g_k(s))''|_{s=\hat{\mu}}}}$; m is the number of quadrature points, t_r are the roots of mth order Hermite polynomial and w_r are corresponding weights [33].

Finding the posterior mode can be computationally cumbersome, calculating $\hat{\sigma}$ requires evaluating the logarithm of g_k at this mode. As we assume that the effects sizes are distributed symmetrically around zero, we use $\hat{\mu}=0$ which avoids numerical posterior mode calculations and evaluating the second derivative at different posterior modes.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Posterior inclusion probability

Combining the previous results we get a numerical solution for calculating the posterior inclusion probability. For every $k > 0$ the inclusion probabilities are

$$p(\gamma_j = k|\Theta) = \frac{\pi_k \int g_k(s)ds}{\pi_0\sqrt{\pi}\exp\{\sum_{i=1}^n[\exp(u_i) - \exp(v_i)]\} + \sum_{l=1}^L \pi_l \int g_l(s)ds} \approx$$
$$\frac{\pi_k\hat{\sigma}_k\sqrt{2}\sum_{r=1}^m w_r g_k(\hat{\sigma}_k\sqrt{2}t_r)}{\pi_0\sqrt{\pi}T + \sum_{l=1}^L \pi_l\hat{\sigma}_l\sqrt{2}\sum_{r=1}^m w_r g_l(\hat{\sigma}_q\sqrt{2}t_r)} = \frac{\pi_k\sqrt{2}\hat{\sigma}_k\sum_{r=1}^m w_r g_k(\hat{\sigma}_k\sqrt{2}t_r)/T}{\pi_0\sqrt{\pi} + \sum_{l=1}^L \pi_l\hat{\sigma}_l\sqrt{2}\sum_{r=1}^m w_r g_l(\hat{\sigma}_l\sqrt{2}t_r)/T}. \quad (40)$$

Similarly we can find the probability of excluding ($\gamma_j = 0$) the marker

$$p(\gamma_j = 0|\Theta) = \frac{\pi_0\sqrt{\pi}}{\pi_0\sqrt{\pi} + \sum_{l=1}^L \pi_l\hat{\sigma}_l\sqrt{2}\sum_{r=1}^m w_r g_l(\hat{\sigma}_l\sqrt{2}t_r)/T}. \quad (41)$$

For both cases $\hat{\sigma}_k$ are calculated as

$$\hat{\sigma}_k = \frac{1}{\sqrt{-(\log g_k(s))''|_{s=0}}} = \frac{1}{\sqrt{2}\sqrt{1 + \alpha^2 C_k\sigma_G^2\sum_{i=1}^n x_{ij}^2(\exp(v_i) - \exp(u_i))}}. \quad (42)$$

For computational purposes we evaluate $\frac{g_k(\hat{\sigma}_k\sqrt{2}t_r)}{T}$ as

$$\frac{g_k(\hat{\sigma}_k\sqrt{2}t_r)}{T} = \exp\{-\alpha\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2}\sum_{i=1}^n d_i x_{ij} +$$
$$\sum_{i=1}^n[\exp(u_i - \alpha x_{ij}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2}) - \exp(u_i) + \exp(v_i) - \exp(v_i - \alpha x_{ij}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2})] - (\hat{\sigma}_k\sqrt{2}t_r)^2\} =$$
$$\exp\{-\alpha\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2}\sum_{i=1}^n d_i x_{ij} + \sum_{i=1}^n[(\exp(v_i) - \exp(u_i))(1 - \exp(-\alpha x_{ij}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2}))] - (\hat{\sigma}_k\sqrt{2}t_r)^2\}. \quad (43)$$

Adaptive rejection sampling

To sample α, μ and β_j, δ_q, we use Adaptive Rejection Sampling, initially outlined by Gilks and Wild [34]. The prerequisite of the method is log-concavity of the sampled density function.

The idea of the method is to build an envelope around the log-density. The lower hull is constructed by evaluating function at some pre-specified abscissae and connecting the evaluation results with linear functions resulting in a piecewise linear lower hull. Upper hull can be constructed either by using tangents evaluated at the prespecified abscissae (Derivative based ARS) or by extending the linear functions obtained in the construction of lower hull (Derivative free ARS [35]). Although derivative based method might result in a more accurate upper hull, thus leading to faster sampling, it would still require evaluating derivatives and thus we employ the derivative free method.

The proposals are sampled from appropriately scaled exponent of upper hull from which it is easier to sample. The sampling proposal will go through tests. If the proposal is not accepted then it will be included in the set of used abscissae to create a more accurate envelope in the next round. Therefore, the method requires specifying the log posterior and at least three initial abscissae. It also requires some abscissae larger and smaller than the posterior mode. To set the abscissae for some parameter θ, we could, for example, choose the abscissae $\{\hat{\theta} - c_\theta, \hat{\theta}, \hat{\theta} + c_\theta\}$, where $\hat{\theta}$ is ideally the posterior mode for θ. c_θ is some positive number that would guarantee that at least one of the proposed abscissae would be larger then posterior mode and one smaller. If $\hat{\theta}$ is the posterior mode, then c_θ choice is arbitrary and smaller c_θ are preferred, potentially decreasing the sampling time.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

In addition, the derivative free method requires specifying the minimum and maximum value of the distribution, an assumption which is often incorrect. In practice, it poses no problems as we can simply set the required minima and maxima to be extreme enough so that the distribution is extremely unlikely to reach those values. To sample intercept μ we set the limits to 2 and 5 which after exponentiation would correspond to 7.39 an 148.4 which we believe each of our posterior means should fit in; for α we set the limits to 0 to 40; for non-zero betas we used the previous beta value $\pm 2\sqrt{C_k \sigma_G^2}$ as minimum and maximum limits for sampling as this can adapt to different mixtures and should still safely retain almost the entire posterior distribution. The Adaptive Rejection Sampling was implemented using C code by Gilks (`http://www1.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/web_page/Welcome.html`, accessed 26.08.2020). In the Supplementary Note we provide a proof of the log-concavity of the functions sampled.

Sampling algorithm

We summarise the serial sampling algorithm in Algorithm 1 along with the specification for the prior distributions and the initialisation of the model parameters. Algorithm 2 summarises the Bulk Synchronous Gibbs sampling for BayesW that extends the Algorithm 1. If the number of workers $T = 1$ and the synchronisation rate $u = 1$ then Algorithm 2 reduces down to Algorithm 1.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Algorithm 1: Serial algorithm for BayesW sampling from the posterior distribution $p(\mu, \alpha, \delta, \gamma, \beta, \pi, \sigma_G^2|D)$. Initialisation and prior specification.

Data: Matrix $\mathbf{X}$ of standardised genotypes, Matrix with covariate data Z, vector of last time without an event $\mathbf{y}$, vector of failure indicators $\mathbf{d}$, prior hyperparameters α_0, κ_0, α_σ, β_σ, σ_μ^2, σ_δ^2, $\mathbf{p}_\mathrm{L}$, iterations I. V_0^j, V_1^j and V_2^j denote the partial sums of the exponentiated residuals (defined in Derivations for the sparse calculations), K is the Euler-Mascheroni constant.

Initial values. Initially, we exclude all the variables from the model, thus the initial $\gamma_j = 0$, $\forall j \in \{1, ..., M\}$. We set $\beta_j = 0$, $\forall j$ and $\delta_q = 0$, $\forall q$. The initial value for α is chosen to be the suitably transformed variance of the log sample. The initial value for μ is the mean of the log sample. The σ_G^2 is initialized as the variance of the log sample divided by the total number of markers M.

Parameters for prior distributions. We set priors weakly informative. Otherwise, if available, prior information could be used. To get weakly informative priors, for α prior, we set parameters to be $\alpha_0 = 0.01$ and $\kappa_0 = 0.01$; for σ_G^2, we set parameters to be $\alpha_\sigma = 1$, $\beta_\sigma = 0.0001$; for μ prior, we set parameter $\sigma_\mu^2 = 100$ and similarly for δ_q we set parameter $\sigma_\delta^2 = 100$. The choice of prior for $\mathbf{p}_\mathrm{L}$ is a vector of ones.

```
Initialise ε = y − μ
for iteration ← 1 to I do
    Add the previous effect to the residual: ε ← ε + μ^old
    Sample μ using ARS;
    Subtract the new effect from the residual: ε ← ε − μ
    Shuffle (covariates);
    foreach covariate q do
        ε ← ε + Z^q δ_q^old
        Sample δ_q using ARS
        ε ← ε − Z^q δ_q
    Sample α using ARS;
    Shuffle (markers);
    Calculate exponentiated residuals: ϵ ← exp(αε − K);
    foreach marker j do
        if β_j^old = 0 then
            Calculate V_0^j, V_1^j, V_2^j;
        if β_j^old ≠ 0 then
            ε ← ε + X^j β_j^old;
            ϵ ← exp(αε − K);
            Calculate V_0^j, V_1^j, V_2^j;
        Sample mixture indicator γ_j;
        if γ_j > 0 then
            sample β_j from the γ_j th conditional distribution using ARS;
            ε ← ε − X^j β_j;
            ϵ ← exp(αε − K);
        if γ_j = 0 then
            set β_j = 0;
    Sample π ;
    Sample σ_G^2.
```

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Algorithm 2: Bulk Synchronous Parallel Gibbs sampling with BayesW. Data, parameter initialisation and prior values are set as in Algorithm 1.

Input : Define T parallel workers (tasks) and synchronisation rate u. Each worker $t \in \{1, ..., T\}$ has a its corresponding vector of marker effects β^t (with number of markers in each $\sim \frac{M}{T}$, M is the total number of markers), indicator values γ^t to update and set of T messages $\Delta\varepsilon_t : N \times 1$, N is the sample size.

Initialise variables;
for *iteration* $\leftarrow 1$ **to** I **do**
  Update μ, δ, α (as in Algorithm 1);
  foreach *subset of size* u **do**
    $\Delta\varepsilon_t \leftarrow 0$, $t \in \{1, ..., T\}$;
    for $t \leftarrow 1$ **to** T *in parallel* **do**
      foreach *column* j *from a subset of size* u *of the columns assigned to worker* t **do**
        if $\beta_j^{old} = 0$ **then**
          Calculate V_0^j, V_1^j, V_2^j based on ϵ;
        if $\beta_j^{old} \neq 0$ **then**
          $\tilde{\varepsilon} \leftarrow \varepsilon + \mathbf{X}^j \beta_j^{old}$;
          $\tilde{\epsilon} \leftarrow \exp(\alpha\tilde{\varepsilon} - K)$;
          Calculate V_0^j, V_1^j, V_2^j based on $\tilde{\epsilon}$;
        Sample mixture indicator γ_j;
        if $\gamma_j > 0$ **then**
          sample β_j from the γ_jth conditional distribution using ARS;
        if $\gamma_j = 0$ **then**
          set $\beta_j = 0$;
        if $\left(\beta_j - \beta_j^{old}\right) \neq 0$ **then**
          $\Delta\varepsilon_t \leftarrow \Delta\varepsilon_t - X^j \left(\beta_j - \beta_j^{old}\right)$;
      Wait until all workers are finished processing their sets of u markers;
    $\varepsilon \leftarrow \varepsilon + \sum_{t=1}^{T} \Delta\varepsilon_t$;
    $\epsilon \leftarrow \exp(\alpha\varepsilon - K)$;
  Update π, σ_G^2 (as in Algorithm 1);

Extension to a grouped mixture of regressions model

Here, we now assume that the SNP marker effects come from Φ of disjoint groups, with a reparametrisation of the model parameters to represent the mean of the logarithm of the phenotype as

$$E(\log Y_i|\mu, \beta, \delta, \alpha) = \mu + \sum_{\varphi=1}^{\Phi} x_i^{\varphi} \beta^{\varphi} + z_i\delta, \tag{44}$$

where there is a single intercept term μ, and a single Weibull shape parameter α, but now x_i^{φ} are the standardised marker values in group φ, β^{φ} are the marker estimates for the corresponding group. Each β_j^{φ} is distributed according to:

$$\beta_j^{\varphi} \sim \pi_0^{\varphi}\boldsymbol{\delta_0} + \pi_1^{\varphi}\mathcal{N}\left(0, C_1^{\varphi}\sigma_{G\varphi}^2\right) + \pi_2^{\varphi}\mathcal{N}\left(0, C_2^{\varphi}\sigma_{G\varphi}^2\right) + \ldots + \pi_{L_\varphi}^{\varphi}\mathcal{N}\left(0, C_L^{\varphi}\sigma_{G\varphi}^2\right) \tag{45}$$

where for each SNP marker group prior probabilities of belonging to each of the mixture distribution k is stored in $L_\varphi + 1$-dimensional vector π^{φ} and these mixture proportions are updated in each iteration $\{\pi_0^{\varphi}, \pi_1^{\varphi}, \ldots, \pi_L^{\varphi}\}$. Each mixture component ($\gamma_j = k \geq 1$) is a normal distribution with zero mean and variance

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

$C_k^{\varphi}\sigma^2_{G\varphi}$, where $\sigma^2_{G\varphi}$ represents the phenotypic variance attributable to markers of group φ and C_k^{φ} is group and mixture specific factor showing the magnitude of variance explained by this specific mixture. Thus the mixture proportions, variance explained by the SNP markers, and mixture constants are all unique and independent across SNP marker groups. The formulation presented here of having an independent variance parameter $\sigma^2_{G\varphi}$ per group of markers, and independent mixture variance components, enables estimation of the amount of phenotypic variance attributable to the group-specific effects and enables differences in the distribution of effects among groups. All of the steps shown in previous paragraphs are still valid and now we are using group specific genetic variances $\sigma^2_{G\varphi}$, prior inclusion probabilities π^{φ} and mixture proportions C_k^{φ}. Furthermore, due to the fact that the model is additive, the sum of group-specific genetic variances represents the total genetic variance $\sigma^2_G = \sum_{\varphi=1}^{\Phi}\sigma^2_{G\varphi}$.

Derivations for the sparse calculations

In order to reduce the number of computations and improve running times we derive a sparse representation of genotypes, given that conditional posterior distributions in our scheme are different, we have to derive different update equations. Suppose ξ_{ij} represents the jth SNP allele count (0, 1 or 2) for the ith individual, and $\bar{\xi}_j, s_j$ represent the mean and standard deviation of the $j-th$ SNP in our sample. In the regular setting we would like to use standardised count values ($x_{ij} = \frac{\xi_{ij}-\bar{\xi}_j}{s_j}$) instead and meanwhile speed up the computations by using the knowledge that x_{ij} can have only three values within a SNP.

There are three equations where we can apply sparsity. Firstly, equation 42 for the $\hat{\sigma}_k$ term (for the jth SNP) in the adaptive Gauss-Hermite quadrature can be expressed as

$$\begin{aligned}
\hat{\sigma}_k &= \frac{1}{\sqrt{2}}\Big[1+\alpha C_k\sigma^2_G\sum_{i=1}^{n}\left(\frac{\xi_{ij}-\bar{\xi}_j}{s_j^2}\right)^2(\exp(v_i)-\exp(u_i))\Big]^{-0.5} = \\
&\frac{1}{\sqrt{2}}\Big[1+\frac{\alpha C_k\sigma^2_G}{s_j^2}\Big(\sum_{i=1}^{n}\xi_{ij}^2(\exp(v_i)-\exp(u_i))-2\bar{\xi}_j\sum_{i=1}^{n}\xi_{ij}(\exp(v_i)-\exp(u_i))+\bar{\xi}_j^2\sum_{i=1}^{n}(\exp(v_i)-\exp(u_i))\Big)\Big]^{-0.5} = \\
&\frac{1}{\sqrt{2}}\Big[1+\frac{\alpha C_k\sigma^2_G}{s_j^2}\Big(\sum_{\xi_{ij}=1}(\exp(v_i)-\exp(u_i))+4\sum_{\xi_{ij}=2}(\exp(v_i)-\exp(u_i)) \\
&-2\bar{\xi}_j\sum_{\xi_{ij}=1}(\exp(v_i)-\exp(u_i))-4\bar{\xi}_j\sum_{\xi_{ij}=2}(\exp(v_i)-\exp(u_i))+\bar{\xi}_j^2\sum_{i=1}^{n}(\exp(v_i)-\exp(u_i))\Big)\Big]^{-0.5} = \\
&\frac{1}{\sqrt{2}}\Big[1+\frac{\alpha C_k\sigma^2_G}{s_j^2}\Big((1-2\bar{\xi}_j)\sum_{\xi_{ij}=1}(\exp(v_i)-\exp(u_i))+4(1-\bar{\xi}_j)\sum_{\xi_{ij}=2}(\exp(v_i)-\exp(u_i))+\bar{\xi}_j^2\sum_{i=1}^{n}(\exp(v_i)-\exp(u_i))\Big)\Big]^{-0.5}.
\end{aligned} \tag{46}$$

We see that s_j and $\bar{\xi}_j$ and the expressions containing these terms can be calculated already beforehand for each SNP j.

Secondly, we can use the knowledge about sparsity to simplify expression 43. More specifically

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

$$\sum_{i=1}^{n}[(\exp(v_i)-\exp(u_i))(1-\exp(-\alpha x_{ij}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2}))] =$$
$$\sum_{i=1}^{n}[(\exp(v_i)-\exp(u_i))] - \exp(\frac{\alpha\bar{\xi}_j}{s_j}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2})\sum_{i=1}^{n}[(\exp(v_i)-\exp(u_i))]\exp(-\frac{\alpha\xi_{ij}}{s_j}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2}) =$$
$$\sum_{i=1}^{n}[(\exp(v_i)-\exp(u_i))] - \exp(\frac{\alpha\bar{\xi}_j}{s_j}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2})\cdot\Big[\sum_{\xi_{ij}=0}(\exp(v_i)-\exp(u_i))+$$
$$\exp(-\frac{\alpha}{s_j}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2})\sum_{\xi_{ij}=1}(\exp(v_i)-\exp(u_i)) + \exp(-\frac{2\alpha}{s_j}\hat{\sigma}_k\sqrt{2}t_r\sqrt{2C_k\sigma_G^2})\sum_{\xi_{ij}=2}(\exp(v_i)-\exp(u_i))\Big]. \tag{47}$$

Thirdly, in expression 22 we can rewrite the transformed residuals as

$$\sum_{i=1}^{n}[\exp(\alpha(\log(a_i)-\mu-x_i\beta-z_i\delta)) - \exp(\alpha(\log(y_i)-\mu-x_i\beta-z_i\delta))] =$$
$$\sum_{i=1}^{n}\Big[\exp(u_i-\alpha(\frac{\xi_{ij}-\bar{\xi}_j}{s_j})\beta_j) - \exp(v_i-\alpha(\frac{\xi_{ij}-\bar{\xi}_j}{s_j})\beta_j)\Big] = \exp\big(\frac{\alpha\beta_j\bar{\xi}_j}{s_j}\big)\cdot$$
$$\Big[-\sum_{\xi_{ij}=0}(\exp(v_i)-\exp(u_i)) - \exp\big(-\frac{\alpha\beta_j}{s_j}\big)\sum_{\xi_{ij}=1}(\exp(v_i)-\exp(u_i)) - \exp\big(-\frac{2\alpha\beta_j}{s_j}\big)\sum_{\xi_{ij}=2}(\exp(v_i)-\exp(u_i))\Big]. \tag{48}$$

For all cases, after each update we need to recalculate the difference $\exp(v_i)-\exp(u_i)$ for each individual i. We notice that it is sufficient to use three sums ($\sum_{\xi_{ij}=\xi}(\exp(v_i)-\exp(u_i))$, $\xi \in \{0,1,2\}$) that we denote as V_0^j, V_1^j, V_2^j which are used in both of the final expressions. Thus, we have eliminated the need to calculate exponents and dot products in expressions 42 and 43, reducing them to a series of sparse summations and making the analysis scale sub-linearly with increasing marker number.

Simulation study

We conducted simulations to analyse the performance of our model under model misspecification, where the phenotypic distribution does not conform to a Weibull distribution, and to different censoring levels in the data. We assessed i) estimation of hyperparameters, ii) false discovery rate, and iii) prediction accuracy. We used $M = 50,000$ uncorrelated markers and $N = 5,000$ individuals for whom we simulated effects on $p = 500$ randomly selected markers, heritability (as defined in the Supplementary Note) was set to be $h^2 = 0.5$. Then, we generated phenotypes from the generalised gamma distribution (see Supplementary Note), retaining the mean and the variance on logarithmic scale and thus fixing the heritability, while varying the θ parameter of the generalised gamma distribution between 0 and 2 (five settings of $\theta = 0, 0.5, 1, 1.5, 2$ with $\theta = 1$ corresponding to a Weibull distribution). For these data sets, we also varied the censoring levels of 0%, 20% and 40% (see Supplementary Note). For each of the censoring and phenotypic distribution combinations 25 replicate phenotypic data sets were created, giving a total of 375 data sets.

To compare our approach with other available methods we analyzed each data set using different approaches: a Cox Lasso [13], a Martingale residuals approach with single-marker ordinary least squares (OLS) regression [11], and Martingale residuals with a Bayesian Dirac mixture of regression model (BayesR) [19]. For each of the 25 simulation replicates, across the five generalised gamma θ parameters, we calculated the correlation between the simulated genetic values and a genetic predictor, created from the regression coefficients obtained from each approach, in an independent data set (same number of markers, same causal

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

markers and same effect sizes, with $N = 1,000$ individuals), with results shown in Figure 1a). Secondly, for all four methods we calculated precision-recall curves for the generalised gamma distributions $\theta \in \{0, 1, 2\}$ and censoring levels 0%, 20% and 40% (Figure 1b, Figure S2). Bayesian models used 5 chains with 1,100 iterations for each chain with a burn-in of 100 and no thinning. The BayesW model was estimated using 11 quadrature points. The hyperparameter for Cox Lasso was estimated using 5-fold cross-validation for each simulated data set separately.

Additionally, for the BayesW model, we analysed each of the 375 data sets, using only a single mixture distribution set to a constant of 0.01, or two mixture distributions with constants 0.01, 0.001. We compared these model formulations by accessing the slope between true and estimated non-zero marker effects (Figure 1c) and the estimated heritability, across the range of generalised gamma distributions (Figure 1d). For each of the 5 generalised gamma θ parameter settings, we also calculated the mean false discovery rate (FDR) levels across the 25 replicate simulations given fixed posterior inclusion probabilities (Figure 1e) for both the single- and two-mixture model formulations (Figure 1f). Finally, we tested the impact of using different number of quadrature points by running the model for the Weibull setting data sets. We varied the number of quadrature points from 3 to 15 across 25 simulation replicates, using one mixture distribution (0.01), and investigated the root mean square error (RMSE: estimated/true) for marker effect estimates within 5 top deciles of the simulated marker effect distribution (Figure S3b).

To test the impact of LD among the markers, we used UK Biobank chromosome 22 imputed genotype data ($M = 194,922$ markers, $N = 20,000$ randomly selected individuals, $p = 2000$ randomly selected causal markers, with heritability $h^2 = 0.5$) and we simulated the phenotypes from Weibull distribution, with 25 simulation replicates. We used this data to compare BayesW to the same other methods described above, by calculating the correlation of simulated genetic value and a genetic predictor in an independent data set (same number of markers, same causal markers and same effect sizes, with $N = 4,000$ individuals). We present these results in Figure S1a. In addition, we used the same genetic data set but varied the censoring levels, to examine the stability of the heritability estimate (Figure S1b). Bayesian analyses used 5 chains with 3000 iterations each and a burn-in of 1000 and thinning of 5. The Cox Lasso model was trained the same way as in the uncorrelated case.

Finally, we ran 10 simulations to check the performance of the BSP Gibbs sampling algorithm in a scenario that would be the closest to the empirical UK Biobank data analysis. We used UK Biobank chromosome 1 of the LD $r^2 < 0.9$ pruned data set which is used also for the full analysis ($M = 230,227$ markers, $N = 25,000$ randomly selected individuals, $p = 2,500$ randomly selected causal SNPs, with heritability $h^2 = 0.5$). The phenotypes were simulated from Weibull distribution and models were run with three variance components (0.0001,0.001,0.01). Models were run by varying the number of tasks (parallelism) between 1, 4, 8, 16 and synchronisation rate (number of markers processed by each task until synchronisation) between 1, 5, 10, 20, 50 (Figure S3a). The scenario of 8 tasks (∼30,000 markers per task) and synchronisation rate of 10 is used in the empirical data analysis.

UK Biobank Data

We restricted our discovery analysis of the UK Biobank to a sample of European-ancestry individuals (N=456,426). To infer ancestry, 488,377 genotyped participants were projected onto the first two genotypic principal components (PC) in 2,504 individuals of the 1,000 Genomes project with known ancestries. Using the obtained PC loadings, we then assigned each participant to the closest population in the 1000 Genomes data: European, African, East-Asian, South-Asian or Admixed. As we wished to contrast the genetic basis of different phenotypes, we then removed closely related individuals as identified in the UK Biobank data release. While we expect that our model can accommodate relatedness similar to other mixed linear model approaches, we wished to compare phenotypes at markers that enter the model due to LD with underlying causal variants, and relatedness leads to the addition of markers within the model to capture the phenotypic covariance of closely related individuals.

We used the imputed autosomal genotype data of the UK Biobank provided as part of the data release. For each individual, we used the genotype probabilities to hard-call the genotypes for variants with an imputation quality score above 0.3. The hard-call-threshold was 0.1, setting the genotypes with probability ≤ 0.9 as missing. From the good quality markers (with missingness less than 5% and p-value for Hardy-Weinberg

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

test larger than 10-6, as determined in the set of unrelated Europeans) were selected those with minor allele frequency (MAF) 0.0025 and rs identifier, in the set of European-ancestry participants, providing a data set of 9,144,511 SNPs, short indels and large structural variants. From this we took the overlap with the Estonian Biobank data to give a final set of 8,433,421 markers. From the UK Biobank European data set, samples were excluded if in the UKB quality control procedures they (i) were identified as extreme heterozygosity or missing genotype outliers; (ii) had a genetically inferred gender that did not match the self-reported gender; (iii) were identified to have putative sex chromosome aneuploidy; (iv) were excluded from kinship inference. Information on individuals who had withdrawn their consent for their data to be used was also removed. These filters resulted in a dataset with 382,466 individuals. We then excluded markers of high LD by conducting LD pruning using a threshold of $r^2 = 0.9$ for a 100kb window leaving us with a final set of 2,975,268 markers. This was done in order to decrease the number of markers that were in extremely high LD and thus giving very little extra information but requiring more than two times the computational resources.

We then selected the recorded measures of for the 382,466 to create the phenotypic data sets for age-at-menopause, age-at-menarche and age-at-diagnosis of HBP, T2D or CAD. For each individual i we created a pair of last known time (logarithmed) without an event Y_i and censoring indicator C_i ($C_i = 1$ if the person had the event at the end of the time period, otherwise $C_i = 0$). If the event had not happened for an individual, then the last time without having the event was defined as the last date of assessment centre visit minus date of birth (only month and year are known, exact date was imputed to 15).

For age-at-menopause we used UKB field 3581 to obtain the time if available. We excluded from the analysis 1) women who had reported of having and later not having had menopause or vice versa, 2) women who said they had menopause but there is no record of the time of menopause (UKB field 2724), 3) women who have had hysterectomy or the information about this is missing (UKB field 3591), 4) women whose menopause is before age 33 or after 65. This left us with a total of $N = 151,472$ women of which $108,120$ had the event and $43,352$ had not had an event by the end of the follow-up. For time-to-menarche we used UKB field 2714 and we excluded all women who had no record for time-to-menarche which left us with a total of $N = 200,493$ women of which all had had the event. For age of diagnosis of HBP we used the UKB field 2966 for and we left out individuals who had the HBP diagnosed but there was no information about the age of diagnosis (UKB field 6150) which left us with a total of $N = 371,878$ individuals of which $95,123$ had the event and $276,755$ had not had an event by the end of the follow-up. For age of diagnosis of T2D we used either the UKB field 2976 or field 20009 or the mean of those two if both were available. We excluded individuals who had indicated self-reported "type 1 diabetes" (code 1222) and had Type 1 Diabetes (ICD code E10) diagnosis; we also excluded individuals who did not have any recorded time for the diagnosis of T2D but they had indicated secondary diagnosis (UKB fields 41202 and 41204) of "non-insulin-dependent diabetes mellitus" (ICD 10 code E11) or self-reported non-cancer illness (UKB field 20002) "type 2 diabetes" (code 1223) or "diabetes" (code 1220). That left us with a total of $N = 372,280$ individuals of which $15,813$ had the event and $356,467$ had not had an event by the end of the follow-up. For age of diagnosis of CAD we used the either the minimum of age at angina diagnosed and age heart attack diagnosed (UKB fields 3627 and 3894) or the minimum age indicated to have either two of diagnoses (codes 1074, 1075) in UKB field 20009 or the mean of those if both were available. We excluded individuals who did not have any information about the time of diagnosis but had following primary or secondary diagnoses: ICD 10 codes I20, I21, I22, I23, I24 or I25; self-reported angina (code 1074) or self-reported heart attack/myocardial infarction (code 1075). That left us with a total of $N = 360,715$ individuals of which $17,452$ had the event and $343,263$ had not had an event by the end of the follow-up.

In the analysis we included covariates of sex, UK Biobank recruitment centre, genotype batch and 20 first principal components of the an LD clumped set of 1.2 million marker data set, calculated using flashPCA, to account for the population stratification in a standard way. We did not include any covariates of age or year of birth because these are directly associated to our phenotypes.

Estonian Biobank Data

The Estonian Biobank cohort is a volunteer-based sample of the Estonian resident adult population. The current number of participants-close to 52000–represents a large proportion, 5%, of the Estonian adult population, making it ideally suited to population-based studies [36]. For the Estonian Biobank Data, 48,088

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

individuals were genotyped on Illumina Global Screening (GSA) (N = 32,594), OmniExpress (N = 8,102), CoreExome (N = 4,903) and Hap370CNV (N = 2,489) arrays. We selected only those from the GSA array and imputed the data set to an Estonian reference, created from the whole genome sequence data of 2,244 participants [37]. From 11,130,313 markers with imputation quality score > 0.3, we selected SNPs that overlapped with the UK Biobank LD pruned data set, resulting in a set of 2,975,268 markers. The phenotypic data was constructed similarly to the phenotypes based on the UK Biobank data for the $N_{Est} = 32,594$ individuals genotyped on GSA array. For time-to-event traits if no event had happened then the time is considered censored and the last known age without the event was used, calculated as the last known date without event minus the date of birth. Because only the year of birth is known, birth date and month were imputed as July 1 for age calculations.

For age-at-menopause we excluded women who had reported having menstruation stopped for other reasons which resulted in 6,434 women who had had menopause and 12,934 women who had not had menopause. For age-at-menarche we excluded women who had not reported the age when the menstruation started which resulted in 18,134. For both age-at-menarche and age-at-menopause if the event had occurred, self-reported age during that event was used.

Initially the cases of CAD, HBP or T2D were identified on the basis of the baseline data collected during the recruitment, where the disease information was either retrieved from medical records or self-reported by the participant. Then, the information was linked with the Estonian Health Insurance database that provided additional information on prevalent cases. To construct the phenotypes for the time-to-diagnosis of CAD, HBP or T2D for the individuals with the corresponding diagnosis we used the age at the first appearance of the respective ICD 10 code that was also used for creating the UK Biobank phenotypes. If the self-reported data about the ICD 10 code has only the information about the year, the date and month was imputed as July 1 and if only the date is missing then the date was imputed as 15. Respective case-control phenotypes for CAD, HBP or T2D were defined 0 if the person had not had an event (censored) and 1 if the person had had an event and these binary indicators were adjusted for age and sex. For the T2D phenotype we excluded individuals with a diagnosis of T1D. For CAD we resulted with 30,015 individuals without the diagnosis and 2579 individuals with a diagnosis, for HBP we resulted with 24,135 individuals without the diagnosis and 8459 individuals with a diagnosis and for T2D we resulted with 30,883 individuals without the diagnosis and 1457 individuals with a diagnosis.

Data Analysis Details

The BayesW model was run on the UK Biobank data without groups and with 20 MAF-LD groups that were defined as MAF quintiles and then quartiles within each of those MAF quintiles split by the LD score. The cut-off points for creating the MAF quintiles were 0.006, 0.013, 0.039, 0.172; the cut-off points for creating LD score quartiles were 2.11, 3.08, 4.51 for the first; 3.20, 4.71, 6.84 for the second; 4.70, 6.89, 9.94 for the third; 7.65, 11.01, 15.70 for the fourth and 10.75, 15.10, 21.14 for the fifth MAF quintile. The model without groups was executed with mixture components (0.00001,0.0001,0.001,0.01) (reflecting that the markers allocated into those mixtures explain the magnitude of 0.001%, 0.01%, 0.1% or 1% of the total genetic variance) and the model with groups was executed with (group specific) mixture components (0.0001,0.001,0.01,0.1). Guided by our simulation study (Figure S3b), we used 25 quadrature points for running each of the models. For the model without groups we used five chains and for the model with groups we used three chains. Each of the chains was run for 10,000 iterations with a thinning of 5 giving us 2,000 samples. We applied a stringent criterion of removing the first half of the chain as burn-in, giving the convergence statistics of Figures SS4, SS5, SS6, SS7. That gave 5,000 samples for the model without groups and 3,000 samples for the model with groups for each of the five traits.

The Bulk Synchronous Parallel Gibbs sampling scheme is implemented by partitioning the markers in equal size chunks assigned to workers (MPI tasks) themselves distributed over compute nodes. For the analyses we used 8 tasks per node; due to the differences in sample size we were using different number of nodes to accommodate the data in memory: for time-to-menopause we used 8 nodes, for time-to-menarche we used 10 nodes and for time-to-diagnosis of CAD, HBP or T2D we used 12 nodes. This resulted in splitting the markers between 64 workers for time-to-menopause, 80 workers for time-to-menarche and 96 workers for time-to-diagnosis of CAD, HBP or T2D. For the last case, the average number of markers assigned to one

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

worker is 30,992. We chose to use a synchronisation rate of 10 meaning the synchronisation between all of the workers was done after sampling 10 markers in each of the workers. Both the choice of maximum number of workers and the synchronisation rate are stringent options considering our simulation study results plotted in Figure S3a.

For testing region-based significance for BayesW we used a Posterior Probability of the Window Variance (PPWV) [25]. PPWV requires first setting a threshold of the proportion of the genetic variance explained. Then, based on the posterior distributions we calculated the probability that a each region explained more than the specified threshold of the proportion of the genetic variance and this quantity is denoted as PPWV. The regions were defined by combining markers into 50kb regions or by mapping the markers to the closest gene in 50kb radius. For the 50kb regions we used thresholds of 1/100,000, 1/10,000 and 1/1,000 of the total genetic variance and as the gene-associated regions can contain greatly more markers we used more stringent thresholds of 1/10,000 and 1/1,000 of the total genetic variance. In addition, to make regions more comparable we fixed an upper bound for the markers that can contribute to a region. For 50kb regions we decided to use at most 125 markers per region and for gene-associated regions we used at most 250 markers per regions.

To do the comparison in terms of discovered regions and prediction accuracy we used the summary statistics from the fastGWA method [26]. Because there were no results for our definition of time-to-CAD or time-to-T2D we used time-to-angina and time-to-heart attack summary statistics for comparison with CAD and time-to-diabetes for comparison with T2D. We called a 50kb region significant if the region contained at least one SNP with a p-value $< 5 \cdot 10^{-8}$. To do the prediction into the Estonian Biobank we only used the markers with p-value $< 5 \cdot 10^{-8}$. We did the predictions only for age-at-menarche and age-at-menopause since the number of significant markers for them is higher.

To do the comparison in terms of predictive accuracy with a competing method we also trained the snpnet method [14, 15] with UK Biobank data and then used the estimates to make prediction into Estonian Biobank. To make the two models comparable, we used exactly the same data sizes for estimating the models on the UK Biobank as were used with the BayesW. For all of the traits we decided to use 95% of the sample size as the training data and the rest as the validation data. This was done in order to minimise the loss in power due to not using the entire sample and 5% of the sample gives a sufficiently large validation set. We ran the snpnet model using 16 threads and allocating 250GB of memory. This was sufficient to find the optimal hyperparameter for the traits of time-to-menopause (22 iterations to find the optimal hyperparameter) and time-to-CAD (21 iterations to find the optimal hyperparameter). However, for the other traits the snpnet procedure ran out of memory and it was decided to use the results from the last available iteration (iteration 28 for time-to-HBP, iteration 35 for time-to-menarche, iteration 27 for time-to-T2D). For the traits for which it was not possible to detect the optimal hyperparameter a sensitivity analysis was done by comparing with the previous iterations. Prediction accuracy was virtually the same between the last available iteration and some iterations before that suggesting that the last available iteration was already providing a hyperparameter close to the optimum.

The prediction based on BayesW into Estonian Biobank $\hat{\boldsymbol{g}}$ was calculated by multiplying $\hat{\boldsymbol{g}} = \mathbf{X}_{Est}\hat{\boldsymbol{\beta}}$, where $\mathbf{X}_{Est}$ is $N_{Est} \times M$ matrix of standardised Estonian genotypes (each column is standardised using the mean and the standard deviation of the Estonian data), $\hat{\boldsymbol{\beta}}$ is the $M \times I$ matrix containing the posterior distributions for M marker effect sizes across I iterations. To calculate the prediction into Estonia we used the BayesW model with groups using 3000 iterations which gave us posterior predictive distributions of the genetic values with 3000 iterations. To create the final predictor, we calculated the mean genetic value for each individual across 3000 iterations. We also created the predictor using the estimates from snpnet by multiplying the standardised Estonian genotype matrix with the vector of snpnet effect size estimates. We evaluated the performance of the two predictors by comparing them to the true phenotype value and calculating R^2 and Harrell's C-statistic [27]. Instead of using the exact phenotypes the martingale residuals from the Cox PH model where the true phenotype was regressed on the gender (if applicable) were used to calculate the R^2. That enables calculating the R^2 value using also the censored individuals. Harrell's C-statistic was calculated from the Cox PH model where the true phenotype was regressed on the predictor and gender (if applicable).

The BayesW calculations have been performed using the facilities of the Scientific IT and Application

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Support Center of EPFL and the Helvetios cluster. All of the post-analysis steps were conducted using R software(version 3.6.1) [38].

Data availability

This project uses UK Biobank data under project 35520. The Estonian Biobank data are available upon request from the cohort authors with appropriate research agreements. Summaries of all posterior distributions obtained are provided in Supplementary data sets. Full posterior distributions of the SNP marker effects sizes for each trait are deposited on Dryad `https://datadryad.org/`

Code availability

Our BayesW model is implemented within the software Hydra, with full open source code available at: `https://github.com/medical-genomics-group/hydra`.

Adaptive rejection sampling C code: `http://www1.maths.leeds.ac.uk/~wally.gilks/adaptive.rejection/arms.method/arms_method.zip`

flashPCA `https://github.com/gabraham/flashpca`

Plink1.90 `https://www.cog-genomics.org/plink2/`

fastGWA database `http://fastgwa.info/ukbimp/phenotypes/`

Computing environment `https://www.epfl.ch/research/facilities/scitas/hardware/helvetios/`

Author contributions

MRR conceived and designed the study. MRR and SEO designed the study with contributions from AK and DTB. SEO, MRR, AK, MP and DTB contributed to the analysis. SEO, MRR, AK and DTB derived the equations and the algorithm. SEO, EJO, DTB coded the algorithm with contributions from AK and MRR. SEO and MRR wrote the paper. ZK, KL, KF, and RM provided study oversight, contributed data and ran computer code for the analysis. All authors approved the final manuscript prior to submission.

Author competing interests

The authors declare no competing interests.

Acknowledgements

This project was funded by an SNSF Eccellenza Grant to MRR (PCEGP3-181181), and by core funding from the Institute of Science and Technology Austria and the University of Lausanne; the work of KF was supported by the grant PUT1665 by the Estonian Research Council. We would like to thank Mike Goddard for comments which greatly improved the work, the participants of the cohort studies, and the Ecole Polytechnique Federal Lausanne (EPFL) SCITAS for their excellent compute resources, their generosity with their time and the kindness of their support.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

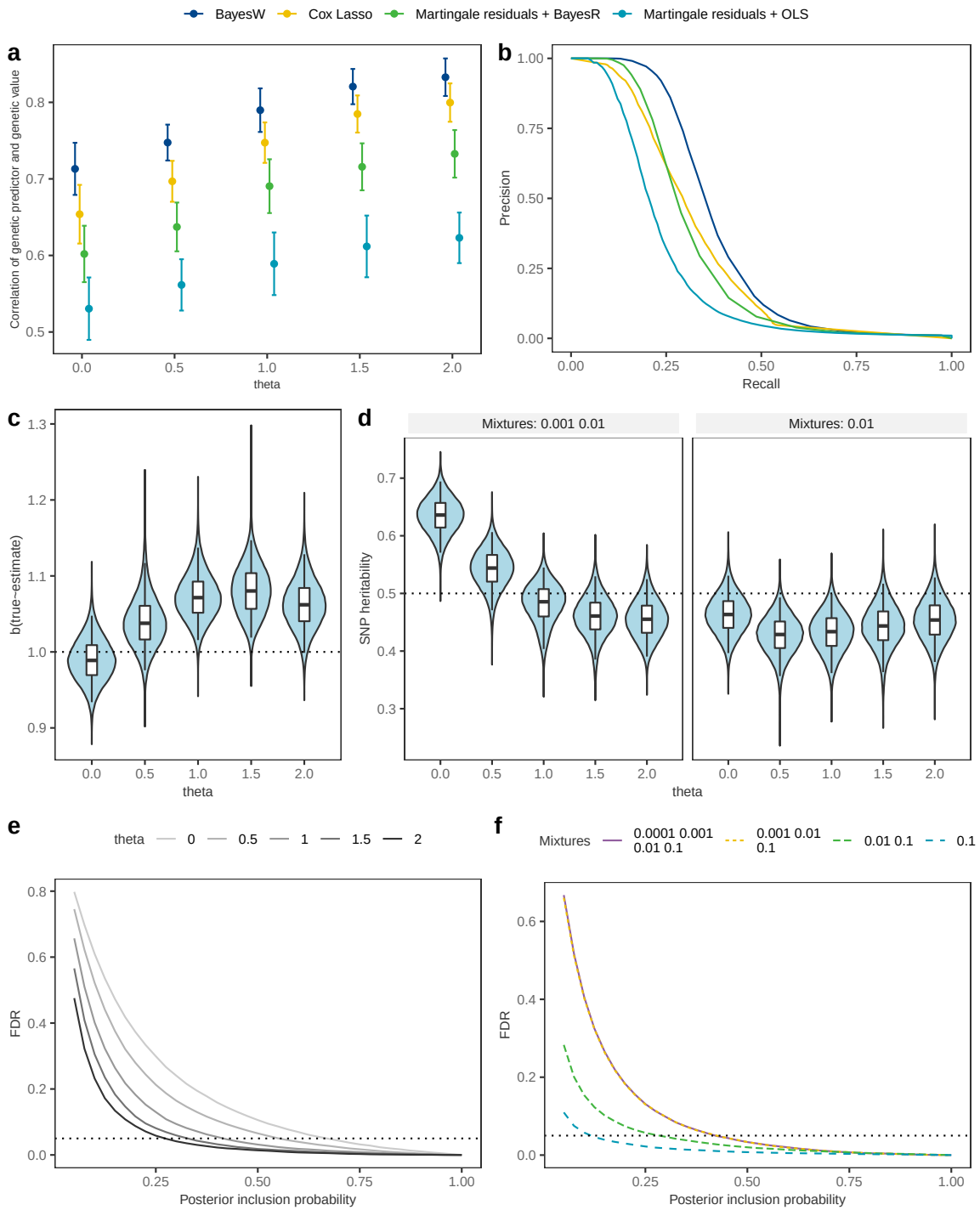

Figure 1. Simulation study results. We present a method comparison, parameter estimation results and the behaviour of the false discovery rate (FDR). Models were estimated on a data set of $M = 50,000$ uncorrelated markers and $N = 5000$ individuals. Phenotypes were created from Generalised gamma distributions (see Supplementary note) using $p = 500$ causals markers and retaining heritability of $h^2 = 0.5$; independent data set had the same markers with $N = 1000$ other individuals. (a) Prediction accuracy of four methods when predicting to an independent data set given different generalised gamma distributions; (b) Precision-recall curves for four methods using Weibull phenotype (theta=1); (c) Regression slope (true effect size $\sim$ estimated effect size) when estimating non-zero marker effects given different theta values estimated with BayesW; (d) BayesW SNP heritability estimates given different generalised gamma distributions and different used mixtures; (e) relationship between the posterior inclusion probability and FDR given different generalised gamma distributions for BayesW; (f) relationship between the posterior inclusion probability and FDR for different number of mixture distributions used using Weibull phenotype and BayesW.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

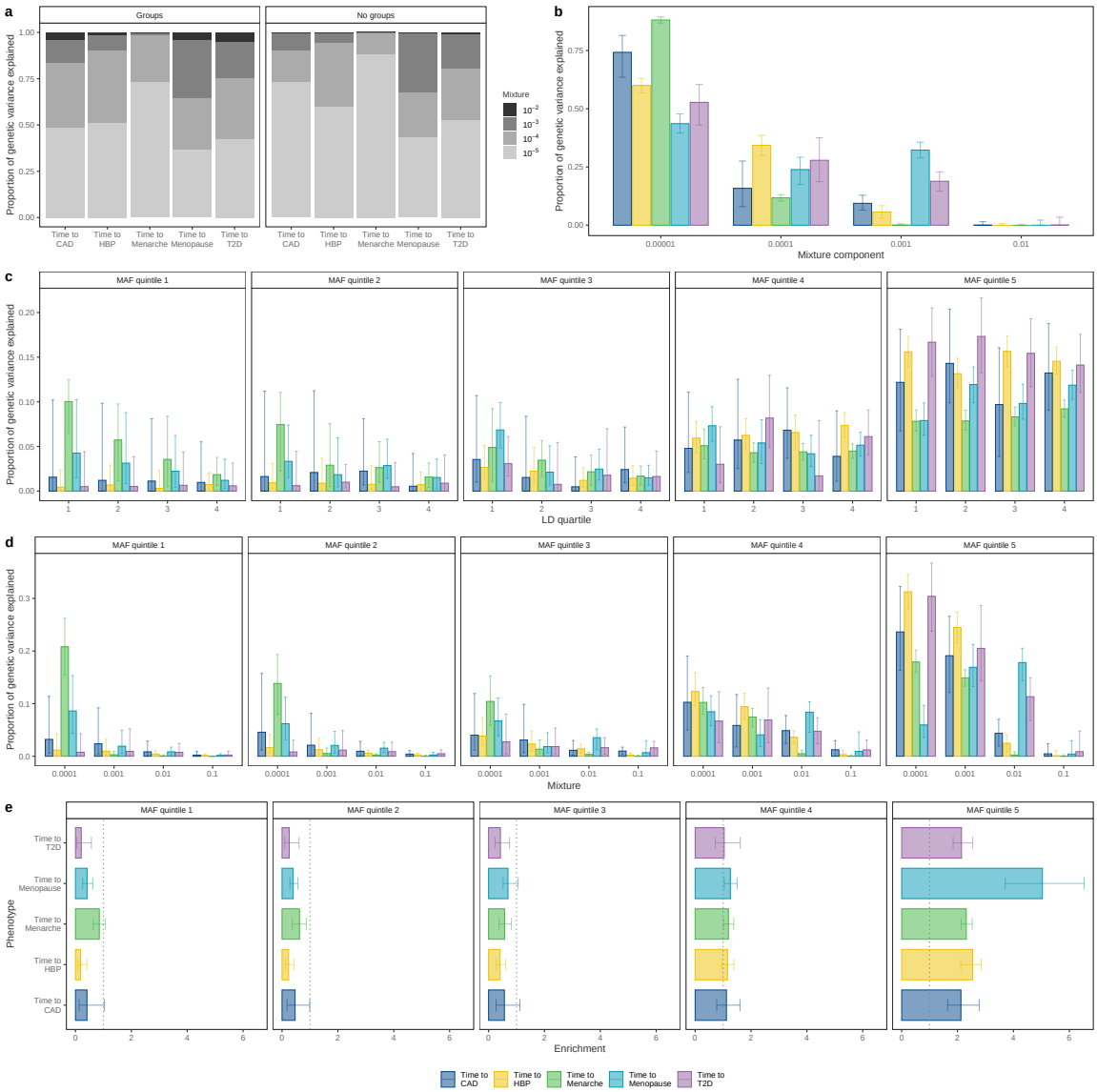

Figure 2. Genetic architecture for time-to-diagnosis of CAD, HBP, T2D and age-at-menarche and age-at-menopause, estimated using 2,975,268 markers and unrelated European ancestry individuals from the UK biobank. We ran BayesW models with and without partitioning the markers to groups. Groups were created by splitting them first to MAF quintiles and each MAF quintile was further split to quartiles based on LD giving us a total of 20 groups. (a) Mean proportions of genetic variances explained by each of the mixtures with the groups model and without groups model, groups model and no group model are yielding rather similar results; (b) distribution of proportion of genetic variance between mixtures for the model without groups, time-to-menarche stands out with almost all of genetic the variance attributed to the small mixtures; (c) distribution of proportion of genetic variance between LD quartiles within each MAF quintile, LD bins do not have a large impact on genetic variance partioning as the credibility intervals are large and medians across LD quartiles are rather stable; (d) distribution of proportion of genetic variance between mixtures within each MAF quintile, mixture allocations tend to be similar compared to no groups model; (e) enrichment (ratio of proportion of genetic variance and proportion of markers attributed to each MAF quintile group) value for each phenotype, enrichment of higher than 1 represents that the markers are explaining more of the genetic variance compared to their count proportion and vice versa. (b)-(e) median and 95% credibility intervals are shown, (c)-(e) are group models. For all of the traits most of the genetic variance is coming from common SNPs (MAF quintile 5).

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

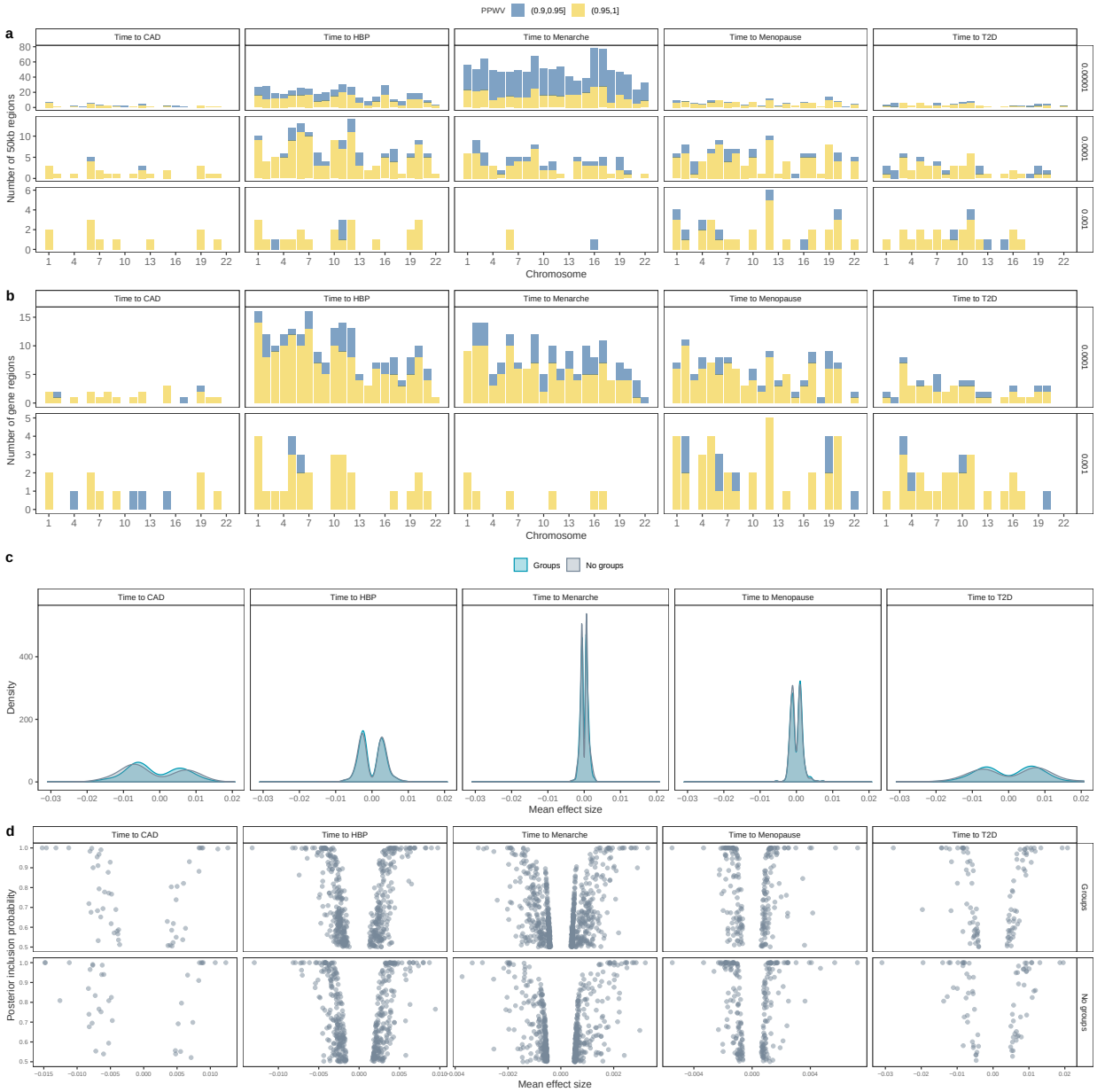

Figure 3. Regional and individual SNP contributions to the time-to-diagnosis of CAD, HBP, T2D and age-at-menarche and age-at-menopause. (a) Count of 50kb regions with high PPWV (Posterior Probability Window Variance). We split the genome into 50kb regions and then we calculated the probability that a region is explaining at least either 0.001%, 0.01% or 0.1% of the genetic variance (PPWV); (b) Count of gene-mapped regions with high PPWV. Each marker is mapped to a closest gene given that it is within ± 50kb from the marker, then for each of those gene-specific regions we calculate PPWV that the region explains either 0.01% or 0.1% of the genetic variance. Both (a) and (b) are using the groups model; (c) distribution of mean non-zero effect sizes for markers with posterior inclusion probability (PIP) > 0.5 for models with and without groups, we pick up mostly large effects for traits such as time-to-diagnosis of CAD or T2D whereas we find an abundance of small effects for age-at-menarche, we see a very small effect of penalisation in the case of group model; (d) relationship between mean non-zero effect size and posterior inclusion probability for markers with PIP > 0.5 with markers.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

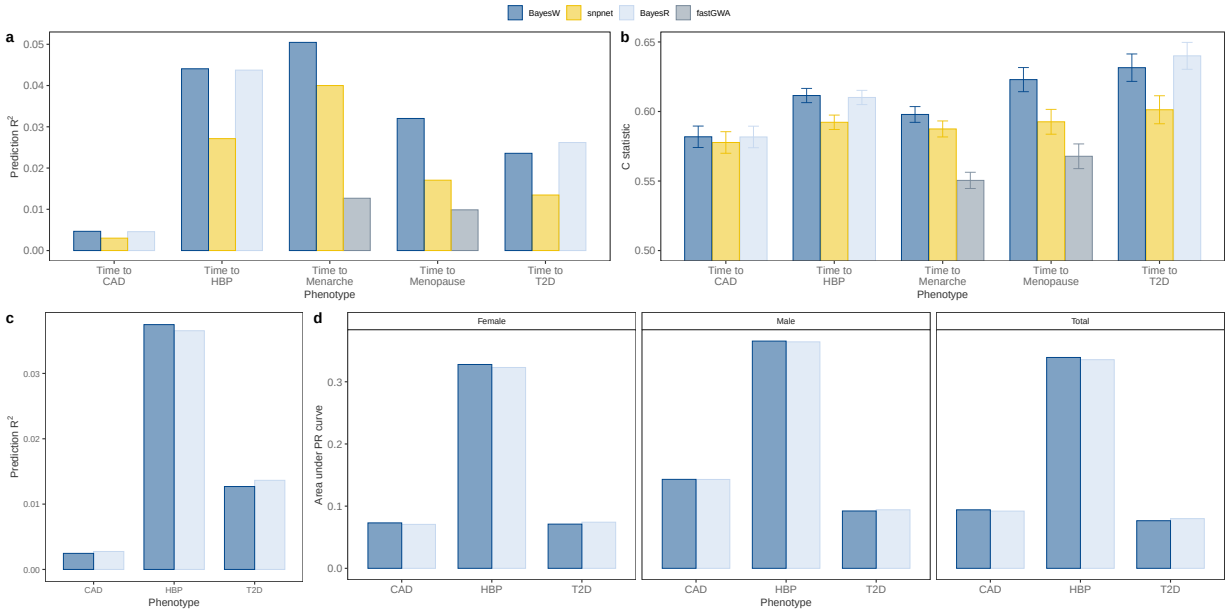

Figure 4. Prediction into the Estonian biobank. BayesW and snpnet methods were used for all of the phenotypes; BayesR was used to see how case-control model predicts time-to-diagnosis phenotypes (CAD, HBP, T2D); fastGWA was used to see how marginal model performs when predicting continuous traits. (a) Prediction R^2 when predicting Estonian martingale residuals of time-to-event phenotypes using either BayesW, snpnet, BayesR or fastGWA model trained on the UK biobank data, martingale residuals were calculated from Cox PH model with an intercept and sex if applicable; (b) Harrell's C-statistic with 95% confidence intervals, calculated from Cox PH model where true phenotype was regressed on the genomic prediction and gender data (for CAD, HBP and T2D); (c) Prediction R^2 when predicting Estonian binary phenotypes (CAD, HBP, T2D) using either model based on time-to-event data (BayesW) or case-control data (BayesR), the binary phenotypes that were used for the comparison were adjusted for age and sex; (d) Area under Precision-recall curve when predicting Estonian binary phenotypes (CAD, HBP, T2D) using either BayesW or BayesR, areas under curve were calculated separately for females, males and everyone combined.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Online Material

Ojavee et al.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

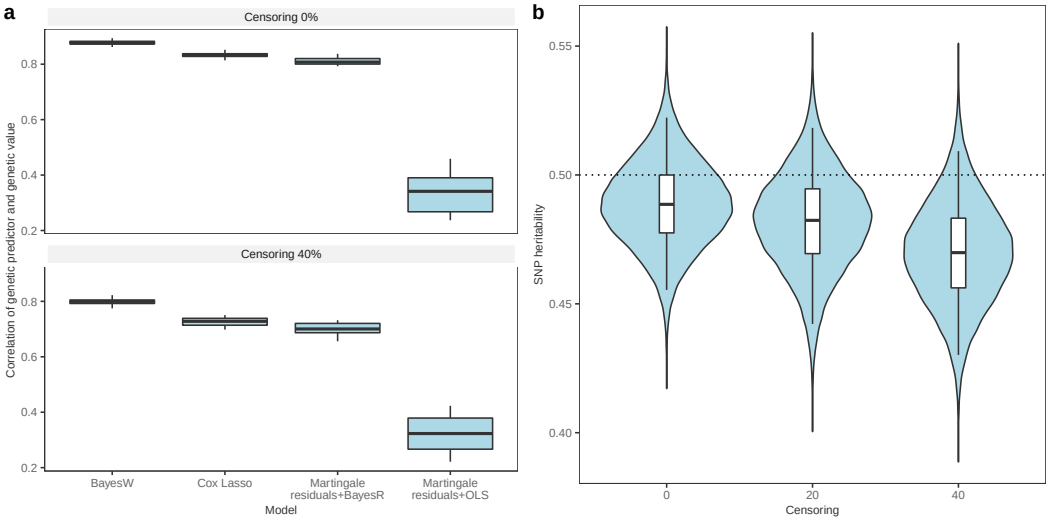

Figure S1. Simulation results for correlated data. 25 simulations used genotypic data from UK biobank randomly chosen individuals and chromosome 22: $N = 20,000$ individuals, $M = 194,922$ markers, $p = 2,000$ causal markers, Weibull phenotype, variance components (0.001,0.01), heritability $h^2 = 0.5$. Independent data set consisted of another 2,000 randomly chosen UK biobank individuals. (a) Prediction accuracy when predicting to an independent data set across four methods given different censoring levels. Similarly to figure 1 we see that also for the correlated markers BayesW gives us higher accuracy for prediction. Higher level of censoring mildly decreases the prediction accuracy; (b) SNP heritability estimates given the censoring level. Similarly to uncorrelated case in figure 1 we see that the true heritability falls into the 95% credibility interval. Higher censoring values mildly decrease the power and therefore also the heritability estimate.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

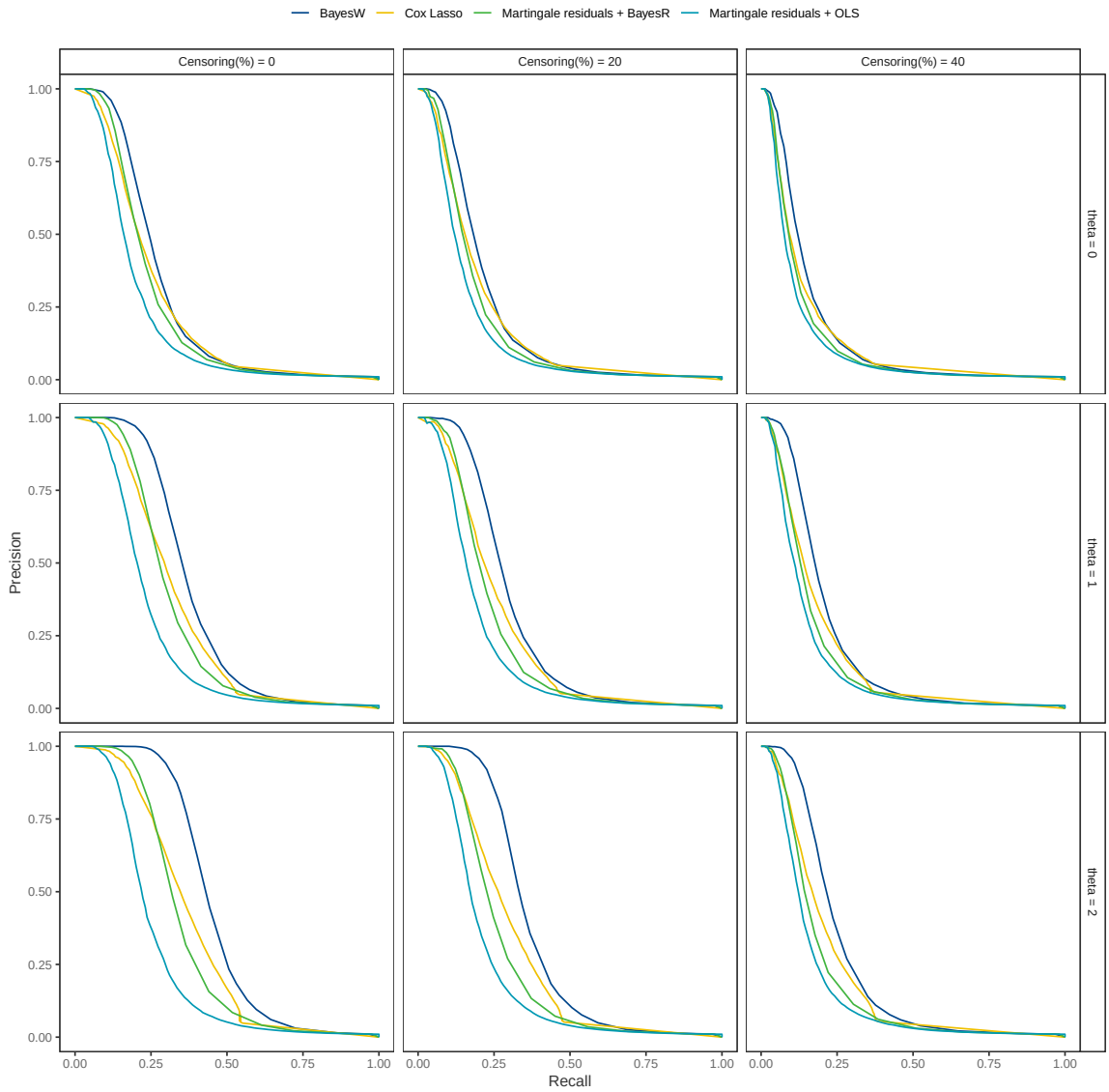

Figure S2. Precision recall curve for different phenotypic distributions and censoring levels. Phenotypes were created by varying the θ parameter of generalised gamma distributions parameter and 25 simulations were run. $\theta = 1$ corresponds to Weibull data and $\theta \to 0+$ (denoted with 0) corresponds to log-normal distribution. Simulation setting: $N = 5,000$ individuals, $M = 50,000$ uncorrelated markers, $p = 500$ causal markers, heritability $h^2 = 0.5$, variance components (0.01). Even for the phenotypes where $\theta \neq 1$, BayesW gets higher precisions for most of the recall values which indicates that the model is relatively robust. The higher the censoring rate the lower is the power.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

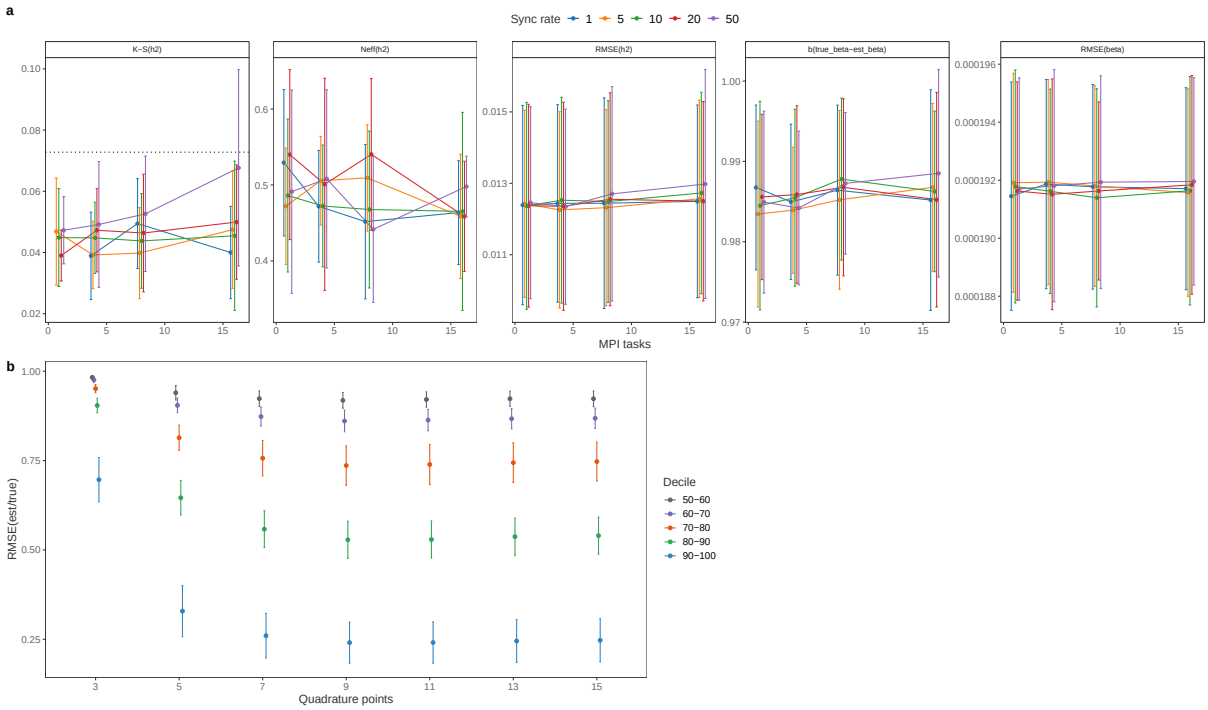

Figure S3. Computational recommendations for using BayesW. (a) Statistics to compare the MPI stability in various scenarios. Compared statistics: 1) Kolmogorov-Smirnov (K-S) test statistic for h^2 where every MPI setting was compared to the sequential version (one task, sync rate = 1), dotted line indicates the critical value for the K-S statistic at significance level of 0.01; 2) Number of effective samples per one iteration for h^2 after thinning of 5; 3) RMSE for h^2; 4) Linear regression slope when regressing the true marker values on the estimated ones; 5) RMSE for β parameters. We took the chromosome 1 of the UK biobank data set that was pruned for LD of 0.9 ($M = 230,227$), with randomly sampled $N = 25,000$ individuals to make it similar to the real data setting. Phenotypes were simulated from Weibull distribution with heritability $h^2 = 0.5$ and the number of causal loci was $p = 2,500$. The models were run with three variance components (0.0001,0.001,0.01) with 10 simulations and 5 chains per MPI setting. The settings with 8 MPI tasks ($\sim 30,000$ markers per task) corresponds roughly to the setting in which full data sets are analysed. In the case where we are using very high synchronisation rate and split markers between many tasks, the estimates might deviate away from the sequential sampling results but for most settings updating markers synchronously will yield similar results compared to sequential sampling. (B) Impact of the number of quadrature points on the estimation of effect sizes. Non-zero effect sizes were grouped to deciles and then RMSE(estimated/true) was calculated within each decile. We see that using more quadrature points makes the estimation better but it the improvement plateaus at some point.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure S4. Convergence diagnostics of model chains for UK Biobank analysis with markers grouped into minor allele frequency (MAF) quintiles and then further subset into linkage disequilibrium (LD) quantiles.(a) traceplot of the proportion of variance attributable to SNP markers across MAF quintiles for each trait , with colours representing the different chains. (b) a time series of the running mean of each chain, of the proportion of variance attributable to SNP markers for each MAF quintile and each trait, showing all chains approach the same mean value for each parameter. (c) lagged autocorrelation plot of each chain, for each MAF quintile and each trait and (d) effective number of uncorrelated sampled obtained for each MAF quintile and each trait. As phenotypic variance is being partitioned it is not expected that posterior estimates obtained are entirely uncorrelated. (e) Geweke z-score statistic comparing the intial part of the chain to the final part, for each MAF quintile and each trait.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

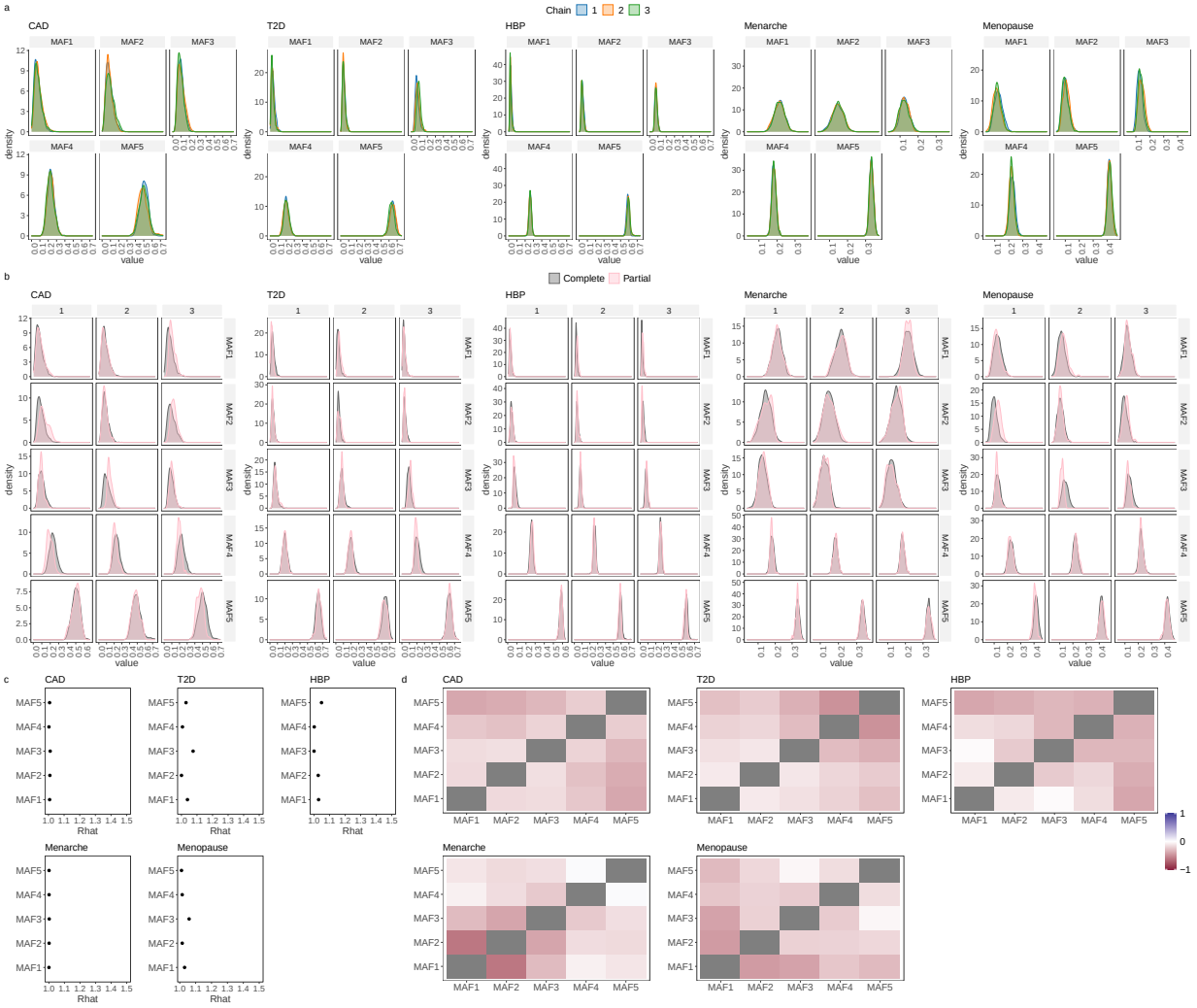

Figure S5. Convergence diagnostics of model chains for UK Biobank analysis with markers grouped into minor allele frequency (MAF) quintiles and then further subset into linkage disequilibrium (LD) quantiles.(a) overlapped density plots to compare the target distribution by chain showing each chain has converged in a similar space, for each MAF quintile and each trait. (b) overlapped density plots comparing the last 10 percent of the chain (green), with the whole chain (pink), showing that the initial and final parts of the chain are sampling the same target distribution for each MAF quintile and each trait. (c) the potential scale reduction factor comparing the among- and within-chain variance for each MAF quintile and each trait. (d) the cross-correlation between all parameters for each MAF quintile and each trait.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

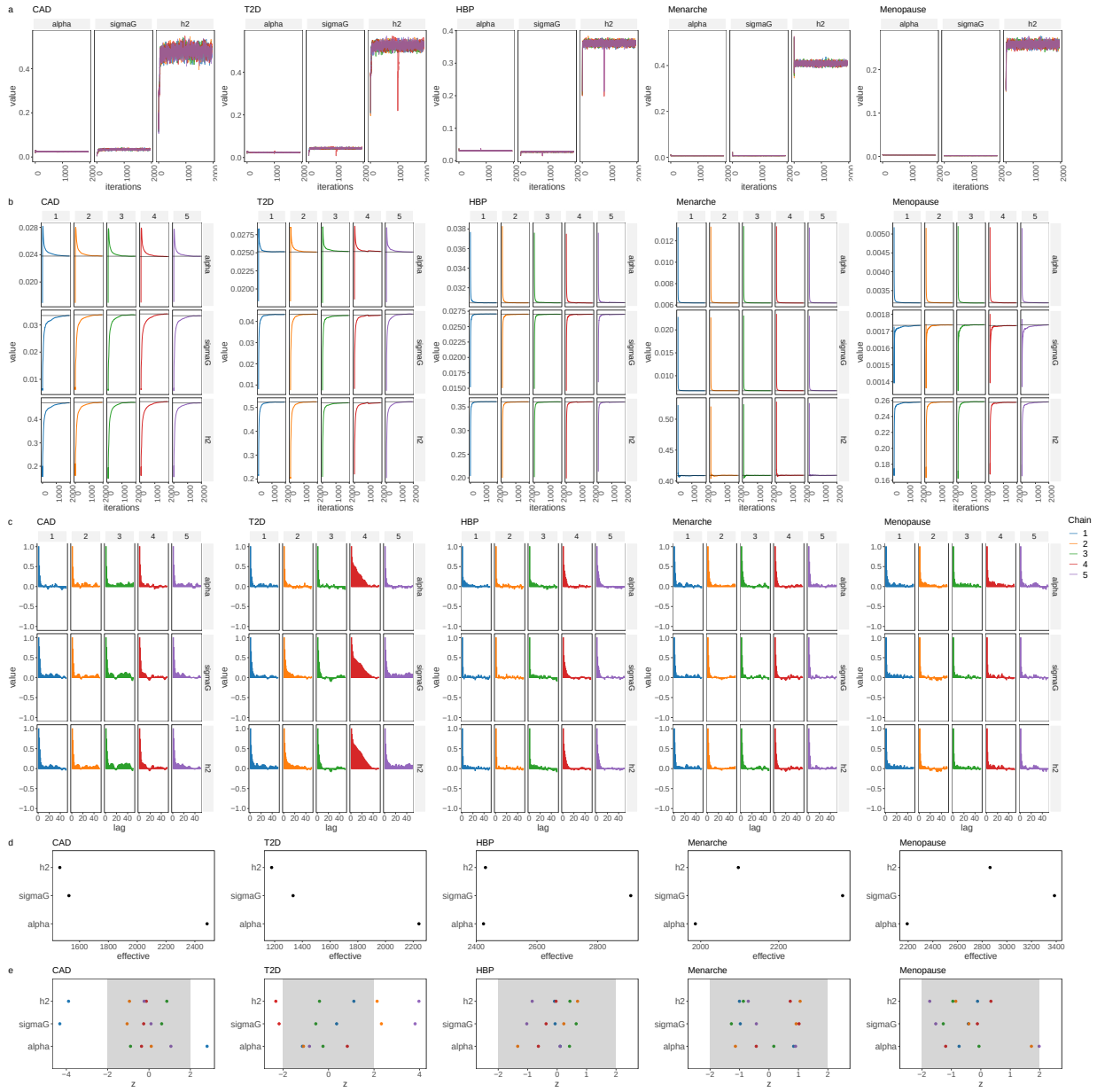

Figure S6. Convergence diagnostics of model chains for UK Biobank analysis with single marker group.(a) traceplot of the residual variance calculated as $\pi^2/(6\alpha^2)$ labelled alpha, phenotypic variance attributable to SNP markers (sigmaG), and the SNP-heritability (h2) of each trait, with colours representing the different chains. (b) a time series of the running mean of each chain, for each trait showing all chains approach the same mean value for each parameter. (c) lagged autocorrelation plot of each chain, for each trait and (d) effective number of uncorrelated sampled obtained for each trait. As phenotypic variance is being partitioned it is not expected that posterior estimates obtained are entirely uncorrelated. (e) Geweke z-score statistic comparing the intial part of the chain to the final part, for each trait.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

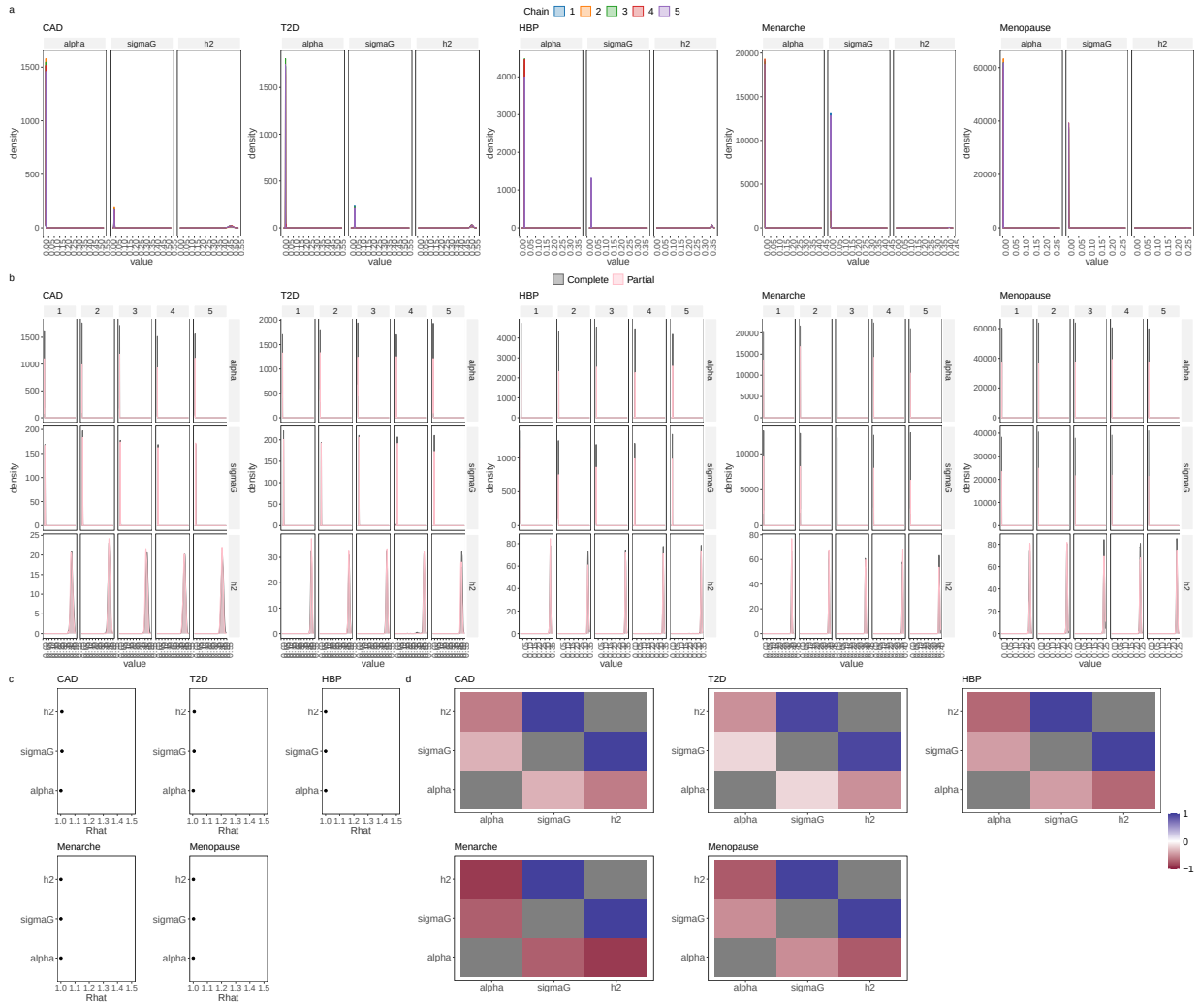

Figure S7. Convergence diagnostics of model chains for UK Biobank analysis with single marker group.(a) overlapped density plots to compare the target distribution by chain showing each chain has converged in a similar space, for each trait, with the residual variance calculated as $\pi^2/(6\alpha^2)$ labelled alpha, phenotypic variance attributable to SNP markers labelled sigmaG), and the SNP-heritability labelled h2. (b) overlapped density plots comparing the last 10 percent of the chain (green), with the whole chain (pink), showing that the initial and final parts of the chain are sampling the same target distribution for each trait. (c) the potential scale reduction factor comparing the among- and within-chain variance for each trait. (d) the cross-correlation between all parameters for each trait.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

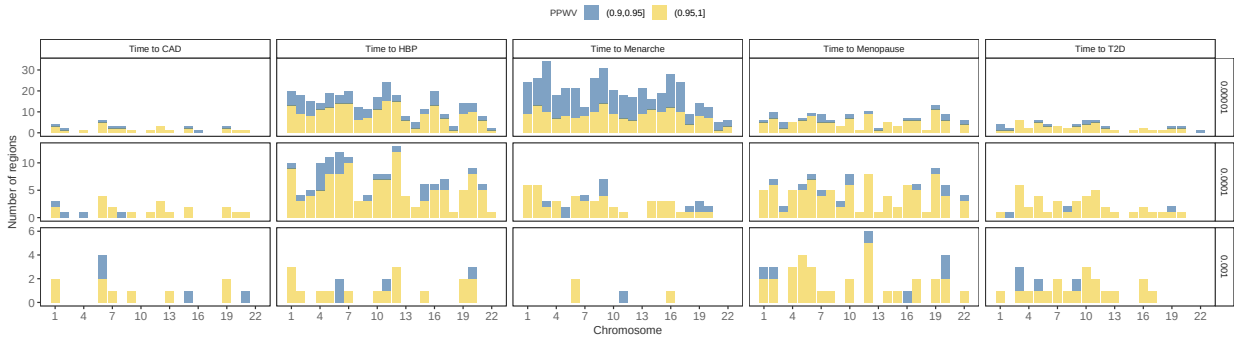

Figure S8. 50kb region contributions to the time-to-diagnosis of CAD, HBP, T2D and age-at-menarche and age-at-menopause using no groups model. Count of 50kb regions with high PPWV (Posterior Probability Window Variance). We split the genome into 50kb regions and then we calculated the probability that a region is explaining at least either 0.001%, 0.01% or 0.1% of the genetic variance (PPWV) using no groups model.

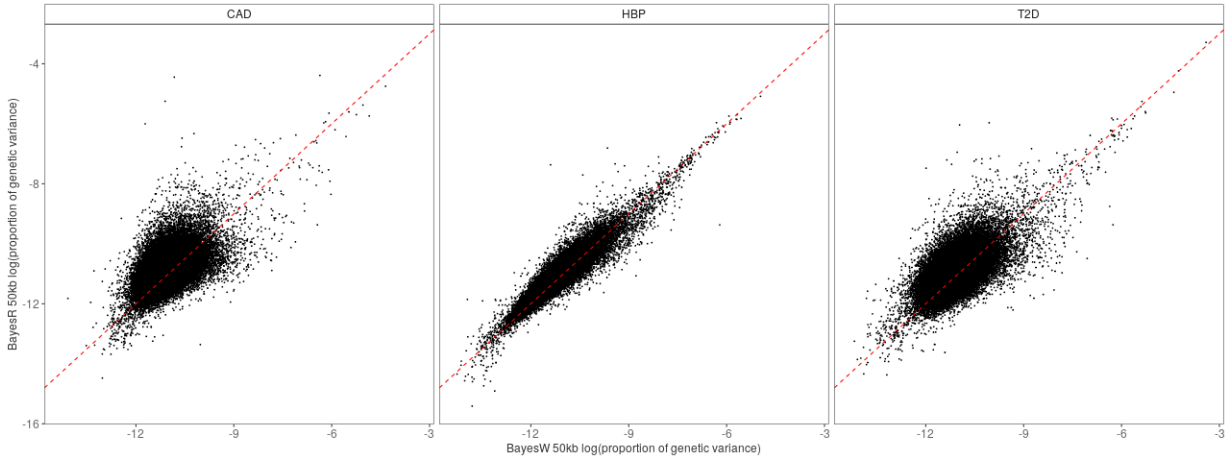

Figure S9. Logarithm of mean proportion of genetic variance explained by each 50kb region using either time-to-event model (BayesW groups) or case-control approach (BayesRR). Three analysed phenotypes were (time-to-)HBP, (time-to-)T2D and (time-to-)CAD and the analysis was conducted on unrelated UK biobank individuals. The BayesW groups model is using 20 MAF-LD groups and BayesR model is using 36 groups based on genomic annotations and MAF-LD binning. Both modelling framework reach similar conclusions in terms of discovered regions: for HBP, T2D and CAD the correlation between the logarithm of results are 0.941, 0.647 and 0.554 respectively.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

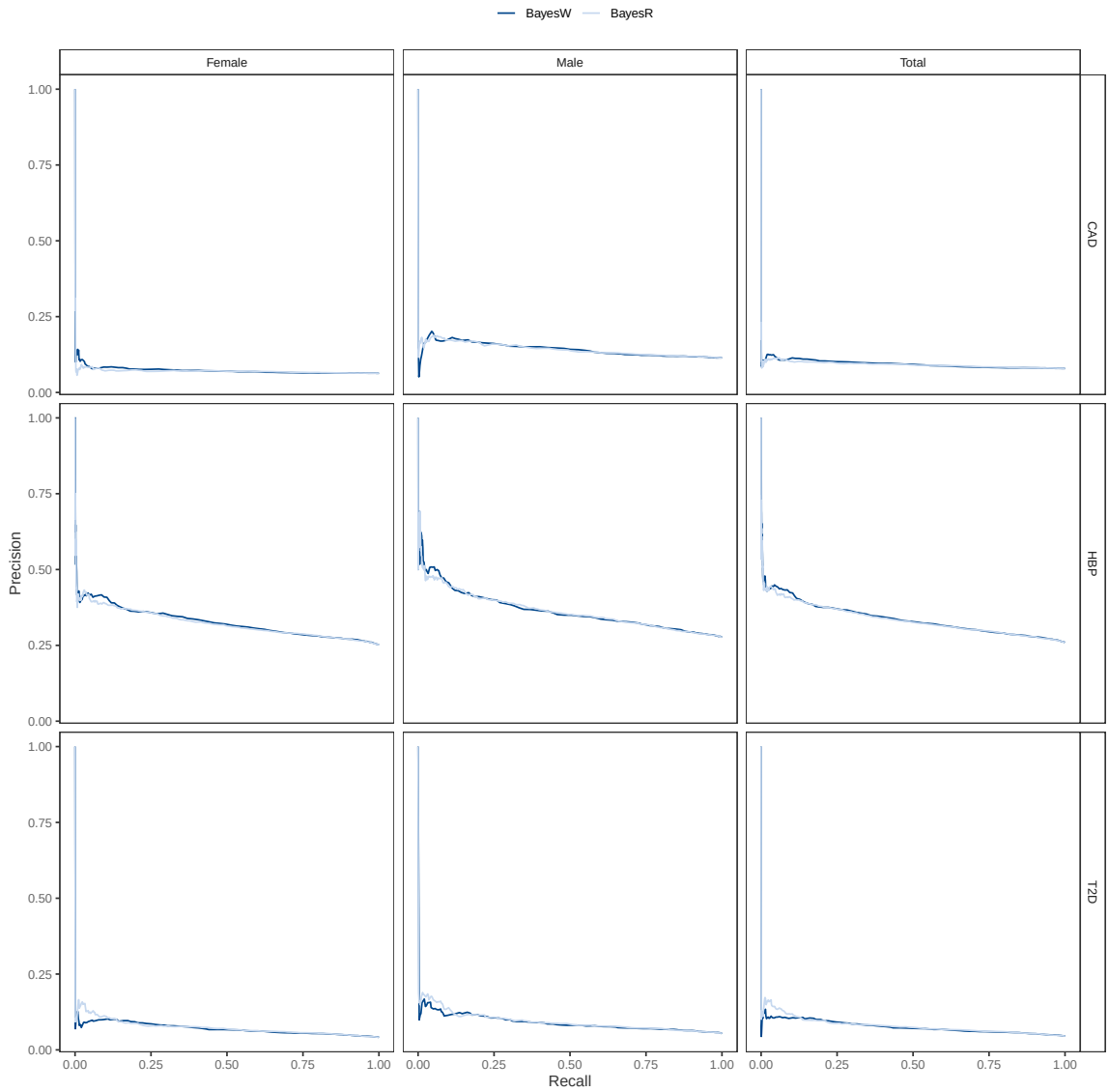

Figure S10. Precision-recall curves for predicting binary phenotypes (CAD, HBP, T2D) in Estonian biobank using either BayesR or BayesW model. Both models were trained on UK biobank data but BayesR model was using binary case-control phenotypes treating them as continuous variables and BayesW model was using respective time-to-diagnosis phenotypes. Precision-recall curves were drawn separately for females, males and all of the people (Total).

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

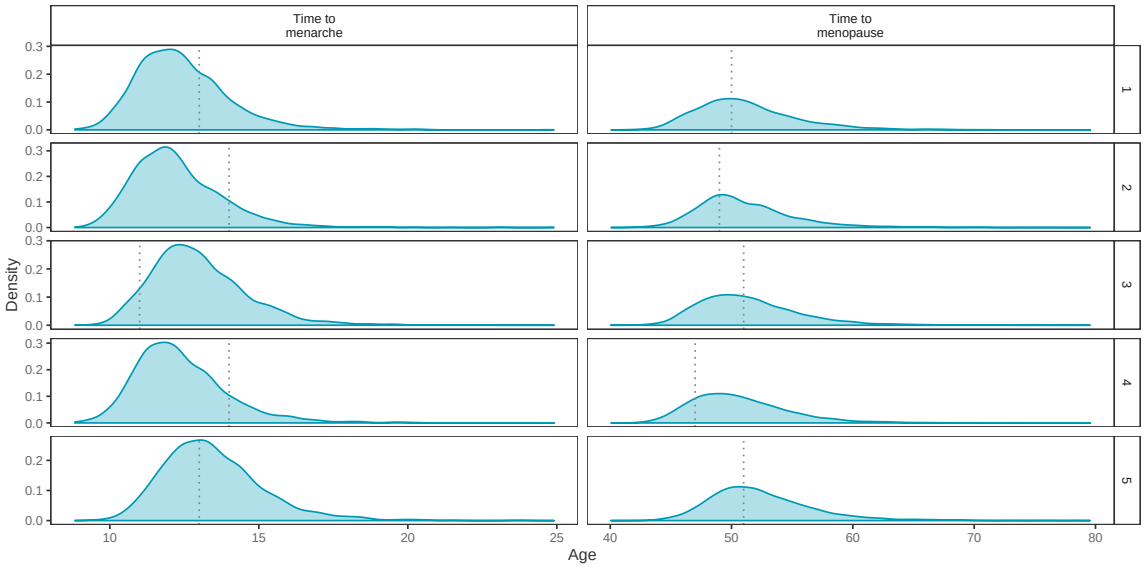

Figure S11. Posterior predictive distributions from BayesW for five random individuals from the Estonian biobank for time-to-menarche and time-to-menopause. BayesW enables calculating individual posterior predictive distributions, dotted lines indicate the true age at event. For age-at-menarche and age-at-menopause 94.8% and 92.3% of the true phenotypes from the Estonian biobank lie within 95% credibility intervals of those predictive distributions.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Supplementary Note

Generating data from Generalised gamma distribution

We generate the time-to-event phenotypes Y_i that follows Generalised gamma distribution using the following formula:

$$\log Y_i = \mu + x_i\beta + cw_i - cEw_i, \tag{49}$$

where w_i (random error) has the following density:

$$f(w) = \frac{|\theta|(\theta^{-2}\exp(\theta w))^{\frac{1}{\theta^2}}\exp(-\theta^{-2}\exp(\theta w))}{\Gamma(\frac{1}{\theta^2})}.$$

This specification means that $E[\log(Y_i|\mu + x_i\beta)] = \mu + x_i\beta$. We will specify the constant c so that the phenotype $\log(Y_i)$ would have a fixed heritability h^2:

$$h^2 = \frac{\sigma_G^2}{\sigma_G^2 + c^2 Var(w_i)} \iff c = \sqrt{\left(\frac{1}{h^2} - 1\right)\frac{\sigma_G^2}{Var(w_i)}}. \tag{50}$$

For each value of θ we calculate the values of $E(w_i)$ and $Var(w_i)$. The value $\theta = 0$ is not a valid parameter for the Generalised gamma distribution. Therefore, the notation $\theta = 0$ stands for the limiting distribution in the process $\theta \to 0$. It is known that the limiting distribution is the log-normal distribution. If $\theta = 1$ then Y_i has a Weibull distribution.

Generating data with fixed proportion of censored individuals

Suppose that time of censoring is C_i for individual i. The observed event is defined as $T_i = \min(Y_i, C_i)$, where Y_i is the true time of event which in reality can be unobserved. The time of censoring is simulated from uniform distribution $C_i \sim U(0, \tau)$. The parameter of the uniform distribution τ is chosen such that the proportion of censored individuals p_τ (that is for whom $C_i < Y_i$) would be some fixed constant. The choice of τ will be dependent on the distribution of Y_i.

Such generative model guarantees that the censoring times C_i and the event times Y_i are independent which is one of the assumptions of our survival model.

Therefore, the τ is chosen such that we would get a fixed censoring level p_τ:

$$p_\tau = E_I[P(C_I < Y_I|I = i)],$$

where I denotes a random individual.

Proofs of log-concavity

The functions g under investigation are twice differentiable. Therefore, to prove the concavity of g it is sufficient to show that $g''(x) \leq 0$ for every x in the domain of g.

Log-concavity for the posterior of α. As constants do not affect concavity, it is sufficient to show that the following function is concave where $\alpha > 0$.

$$g(\alpha) = \log p(\alpha|D, \mu, \beta) - const =$$

$$(\alpha_0+d-1)\log\alpha+\alpha[\sum_{i=1}^{n} d_i(\log y_i-\mu-x_i\beta)-\kappa_0]+e^{-K}\sum_{i=1}^{n}[\exp(\alpha(\log(a_i)-\mu-x_i\beta))-\exp(\alpha(\log(y_i)-\mu-x_i\beta))] =$$

$$A\log\alpha + B\alpha + e^{-K}\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta))] =$$

$$A\log\alpha + B\alpha + e^{-K}\sum_{i=1}^{n}[\exp(\alpha C_i) - \exp(\alpha D_i)]$$

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We see that $A > 0$ and as failure or censoring happens after left truncation then $D_i > C_i$ for every i. The second derivative of g is

$$g''(\alpha) = -\frac{A}{\alpha^2} + e^{-K}\sum_{i=1}^{n}[C_i^2 \exp(\alpha C_i) - D_i^2 \exp(\alpha D_i)].$$

As $A > 0$ and $D_i > C_i$ for every i, then $g''(\alpha) < 0$ for every $\alpha > 0$.

Log-concavity for the posterior of β_j. We need to show that the following function is concave

$$g(\beta_j) = \log p(\beta_j|D, \alpha, \mu, \beta_{-j}) - const =$$
$$-\alpha\beta_j \sum_{i=1}^{n} d_i x_{ij} + \exp(-K)\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta))] - \frac{1}{2\sigma_\beta^2}\beta_j^2 =$$
$$-A\beta_j - B\beta_j^2 + \sum_{i=1}^{n}(C_i - D_i)\exp\{-\alpha x_{ij}\beta_j\}.$$

We see that $B > 0$, and $D_i > C_i$, $i \in \{1, ..., n\}$. The second derivative is

$$g''(\beta_j) = -B + \alpha^2 \sum_{i=1}^{n}(C_i - D_i)x_{ij}^2 \exp(-\alpha x_{ij}\beta_j)$$

and clearly $g''(\beta_j) < 0$ for every $\beta_j \in \mathbb{R}$.

Log-concavity for the posterior of δ_q. Analogous to the case of β_j.

Log-concavity for the posterior of μ.

$$g(\mu) = \log p(\mu|D, \alpha, \beta) - const = -\alpha\mu d +$$
$$\exp(-K)\sum_{i=1}^{n}[\exp(\alpha(\log(a_i) - \mu - x_i\beta)) - \exp(\alpha(\log(y_i) - \mu - x_i\beta))] - \frac{1}{2\sigma_\mu^2}\mu^2 =$$
$$-A\mu - B\mu^2 + \sum_{i=1}^{n}(C_i - D_i)\exp\{-\alpha\mu\}$$

We see that $B > 0$, and $D_i > C_i$, $i \in \{1, ..., n\}$. The second derivative is

$$g''(\mu) = -B + \alpha^2 \sum_{i=1}^{n}(C_i - D_i)\exp(-\alpha\mu)$$

and clearly $g''(\mu) < 0$ for every $\mu \in \mathbb{R}$.

SNP heritability of age-at-onset on the log and original scale

We will derive expressions for SNP heritability on the log-scale and on the original scale given that the phenotype follows a Weibull distribution. The quantity defined here is meaningful in terms of heritability if the appearance of the event is guaranteed for all of the individuals (for example menopause but not diagnosis of T2D). For the events that are not guaranteed to happen we call this quantity pseudo-heritability because then the underlying random variable is improper.

The Weibull data for the model is generated using following expression

$$\log Y_i = \mu + x_i\beta + \frac{w_i}{\alpha} - \frac{K}{\alpha}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

where w_i (random error) comes from the standard extreme value distribution (Gumbel distribution) and K is Euler-Mascheroni constant. This guarantees that Y_i has a Weibull distribution, $E(\log(Y_i)|\mu + x_i\beta) = \mu + x_i\beta$ and $Var(\log(Y_i)|\mu + x_i\beta) = \frac{\pi^2}{6\alpha^2}$. In the following, we denote $g_i = x_i\beta$ and we assume $g_i \sim N(0, \sigma_G^2)$ where σ_G^2 is the genetic variance of the logarithmed phenotype.

We require the variance of $\log Y_i$ and the corresponding heritability $h^2_{\log}$. Using the law of total variance it is possible to separate the variance components in the following way where the first part represents the genetic variance and second is the error variance.

$$Var(\log Y_i) = E[Var(\mu + g_i + \frac{w_i}{\alpha} - \frac{K}{\alpha}|w_i)] + Var[E(\mu + g_i + \frac{w_i}{\alpha} - \frac{K}{\alpha}|w_i)] = \sigma_G^2 + Var(\mu - \frac{K}{\alpha} + \frac{w_i}{\alpha}) = \sigma_G^2 + \frac{\pi^2}{6\alpha^2} \quad (51)$$

As the genetic variance is σ_G^2 we get the log-scale heritability by dividing the genetic variance component by the total variance:

$$h^2_{\log} = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\pi^2}{6\alpha^2}}. \quad (52)$$

Using the same idea we decompose the variance of Y_i and find the corresponding heritability h^2_{orig}. In addition we need three following results:

Firstly we see that $[\exp(\mu - \frac{K}{\alpha})]^2$ will cancel out when calculating the heritability

$$Var(Y_i) = Var[\exp(\mu + g_i + \frac{w_i}{\alpha} - \frac{K}{\alpha})] = [\exp(\mu - \frac{K}{\alpha})]^2 Var[\exp(g_i + \frac{w_i}{\alpha})]. \quad (53)$$

Secondly, we note that as g_i is normally distributed then $\exp(g_i)$ has a corresponding log-normal distribution and therefore $Var(\exp(g_i)) = [\exp(\sigma_g^2) - 1]\exp(\sigma_g^2)$ and $E(\exp(g_i)) = \exp(\sigma_G^2/2)$.

Thirdly, we see that if w_i has standard extreme value distribution then for $\alpha > 2$

$$E[\exp(\frac{2w_i}{\alpha})] = \Gamma(1 - \frac{2}{\alpha}),$$

otherwise the expected value is undefined and for $\alpha > 1$

$$E[\exp(\frac{w_i}{\alpha})] = \Gamma(1 - \frac{1}{\alpha}),$$

otherwise the expected value is undefined.

Using the law of total variance the genetic and error variance can be separated as

$$Var[\exp(g_i + \frac{w_i}{\alpha})] = E[Var(\exp(g_i + \frac{w_i}{\alpha})|w_i)] + Var[E(\exp(g_i + \frac{w_i}{\alpha})|w_i)]. \quad (54)$$

The genetic variance component is

$$E[Var(\exp(g_i + \frac{w_i}{\alpha})|w_i)] = E[(\exp(\frac{w_i}{\alpha}))^2 Var(\exp(g_i))] = E[\exp(\frac{2w_i}{\alpha})]Var(\exp(g_i)) = \Gamma(1 - \frac{2}{\alpha})[\exp(\sigma_G^2) - 1]\exp(\sigma_G^2). \quad (55)$$

The error variance component is

$$Var[E(\exp(g_i + \frac{w_i}{\alpha})|w_i)] = Var[\exp(\frac{w_i}{\alpha})E(\exp(g_i))] = \left(\exp\left(\frac{\sigma_g^2}{2}\right)\right)^2 Var\left(\exp\left(\frac{w_i}{\alpha}\right)\right) = \exp(\sigma_G^2)\left(\Gamma(1 - \frac{2}{\alpha}) - \left(\Gamma(1 - \frac{1}{\alpha})\right)^2\right). \quad (56)$$

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Therefore, by dividing the genetic variance component with the sum of genetic and error variance, we get that the heritability on the original scale, given $\alpha > 2$, is

$$h^2_{orig} = \frac{\Gamma(1-\frac{2}{\alpha})(\exp(\sigma^2_G)-1)}{\Gamma(1-\frac{2}{\alpha})\exp(\sigma^2_G) - \left(\Gamma(1-\frac{1}{\alpha})\right)^2}. \tag{57}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 years of gwas discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5 – 22, 2017.

2. James R. Staley, Edmund Jones, Stephen Kaptoge, Adam S. Butterworth, Michael J. Sweeting, Angela M. Wood, Joanna M. M. Howson, and on behalf of the EPIC-CVD Consortium. A comparison of cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. *European Journal of Human Genetics*, 25(7):854–862, Jul 2017.

3. Hamzah Syed, Andrea L Jorgensen, and Andrew P Morris. Evaluation of methodology for the analysis of 'time-to-event' data in pharmacogenomic genome-wide association studies. *Pharmacogenomics*, 17(8):907–915, 2016. PMID: 27248145.

4. Kristi Läll, Reedik Mägi, Andrew Morris, Andres Metspalu, and Krista Fischer. Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores. *Genetics in Medicine*, 19(3):322–329, Mar 2017.

5. David. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

6. Hamzah Syed, Andrea L. Jorgensen, and Andrew P. Morris. Survivalgwas_sv: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. *BMC Bioinformatics*, 18(1):265, 2017.

7. Hamzah Syed, Andrea L. Jorgensen, and Andrew P. Morris. Survivalgwas_power: a user friendly tool for power calculations in pharmacogenetic studies with "time to event" outcomes. *BMC bioinformatics*, 17(1):523–523, Dec 2016. 27931206[pmid].

8. Abbas A Rizvi, Ezgi Karaesmen, Martin Morgan, Leah Preus, Junke Wang, Michael Sovic, ..., and Lara E Sucheston-Campbell. gwasurvivr: an R package for genome-wide survival analysis. *Bioinformatics*, 35(11):1968–1970, 11 2018.

9. Wenjian Bi, Lars G. Fritsche, Bhramar Mukherjee, Sehee Kim, and Seunggeun Lee. A fast and accurate method for genome-wide time-to-event data analysis and its application to uk biobank. *The American Journal of Human Genetics*, 107(2):222 – 233, 2020.

10. Peter K. Joshi, Krista Fischer, Katharina E. Schraut, Harry Campbell, Tõnu Esko, and James F. Wilson. Variants near chrna3/5 and apoe have age- and sex-related effects on human lifespan. *Nature Communications*, 7(1):11174, 2016.

11. Peter K. Joshi, Nicola Pirastu, Katherine A. Kentistou, Krista Fischer, Edith Hofer, Katharina E. Schraut, ..., and James F. Wilson. Genome-wide meta-analysis associates hla-dqa1/drb1 and lpa and lifestyle factors with human longevity. *Nature Communications*, 8(1):910, 2017.

12. Liang He and Alexander M. Kulminski. Fast algorithms for conducting large-scale gwas of age-at-onset traits using cox mixed-effects models. *Genetics*, 215(1):41–58, 2020.

13. Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

14. Junyang Qian, Yosuke Tanigawa, Wenfei Du, Matthew Aguirre, Chris Chang, Robert Tibshirani, Manuel A. Rivas, and Trevor Hastie. A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the uk biobank. *bioRxiv*, 2020.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

15. Ruilin Li, Christopher Chang, Johanne Marie Justesen, Yosuke Tanigawa, Junyang Qian, Trevor Hastie, Manuel A. Rivas, and Robert Tibshirani. Fast lasso method for large-scale and ultrahigh-dimensional cox model with applications to uk biobank. *bioRxiv*, 2020.

16. Paul J. Newcombe, Hamid Raza Ali, Fiona M. Blows, Elena Provenzano, Paul D. Pharoah, Carlos Caldas, and Sylvia Richardson. Weibull regression with bayesian variable selection to identify prognostic tumour markers of breast cancer survival. *Statistical Methods in Medical Research*, 26(1):414–436, 2017.

17. Weiwei Duan, Ruyang Zhang, Yang Zhao, Sipeng Shen, Yongyue Wei, Feng Chen, and David C. Christiani. Bayesian variable selection for parametric survival model with applications to cancer omics data. *Human genomics*, 2018.

18. Leonhard Held, Isaac Gravestock, and Daniel Sabanés Bové. Objective bayesian model selection for cox regression. *Statistics in Medicine*, 35(29):5376–5390, 2016.

19. Marion Patxot, Daniel Trejo Banos, Athanasios Kousathanas, Etienne J. Orliac, Sven E. Ojavee, Gerhard Moser, Julia Sidorenko, Zoltan Kutalik, Reedik Mägi, Peter M. Visscher, Lars Rönnegård, and Matthew R. Robinson. Probabilistic inference of the genetic basis of complex traits at biobank scale. Unpublished, 2020.

20. Luke M. Evans, Rasool Tahmasbi, Scott I. Vrieze, Gonçalo R. Abecasis, Sayantan Das, Steven Gazal, Douglas W. Bjelland, Teresa R. De Candia, Michael E. Goddard, Benjamin M. Neale, et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nature Genetics*, 50(5):737–745, 2018.

21. Doug Speed, Na Cai, Michael R. Johnson, Sergey Nejentsev, David J. Balding, UCLEB Consortium, et al. Reevaluation of snp heritability in complex human traits. *Nature Genetics*, 49(7):986, 2017.

22. Doug Speed, John Holmes, and David J. Balding. Evaluating and improving heritability models using summary statistics. *Nature Genetics*, 52(4):458–462, 2020.

23. Kangcheng Hou, Kathryn S Burch, Arunabha Majumdar, Huwenbo Shi, Nicholas Mancuso, Yue Wu, Sriram Sankararaman, and Bogdan Pasaniuc. Accurate estimation of snp-heritability from biobank-scale data irrespective of genetic architecture. *Nature Genetics*, page 1, 2019.

24. Yuval B. Simons, Kevin Bullaughey, Richard R. Hudson, and Guy Sella. A population genetic interpretation of gwas findings for human quantitative traits. *PLOS Biology*, 16(3):1–20, 03 2018.

25. Rohan Fernando, Ali Toosi, Anna Wolc, Dorian Garrick, and Jack Dekkers. Application of whole-genome prediction methods for genome-wide association studies: A bayesian approach. *Journal of Agricultural, Biological and Environmental Statistics*, 22(2):172–193, Jun 2017.

26. Longda Jiang, Zhili Zheng, Ting Qi, Kathryn E. Kemper, Naomi R. Wray, Peter M. Visscher, and Jian Yang. A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, 51(12):1749–1755, Dec 2019.

27. Frank E. Harrell Jr., Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996.

28. Vincent Ducrocq and George Casella. A bayesian analysis of mixed survival models. *Genetics Selection Evolution*, 28, 12 1996.

29. Inge R. Korsgaard, Per Madsen, and Just Jensen. Bayesian inference in the semiparametric log normal frailty model using gibbs sampling. *Genetics, Selection, Evolution : GSE*, 30(3):241–256, May 1998. PMC2707404[pmcid].

medRxiv preprint doi: https://doi.org/10.1101/2020.09.04.20188441; this version posted September 7, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

30. Vincent Ducrocq. Two year of experience with the franch genetic evaluation od dairy bulls on production-adjusted longevity of their daughters. *Interbull Bulletin*, 21, 01 1999.

31. M. H. Yazdi, P. M. Visscher, V. Ducrocq, and R. Thompson. Heritability, reliability of genetic evaluations and response to selection in proportional hazard models. *Journal of Dairy Science*, 85(6):1563 – 1577, 2002.

32. Edward I. George and Robert E. McCulloch. Approaches for bayesian variable selection. *Statistica Sinica*, 7(2):339–373, 1997.

33. Qing Liu and Donald A. Pierce. A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629, 1994.

34. W. R. Gilks and P. Wild. Adaptive Rejection Sampling for Gibbs Sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992.

35. W. R. Gilks. Derivative-free adaptive rejection sampling for Gibbs sampling. *Bayesian Statistics 4, (eds. Bernardo, J., Berger, J., Dawid, A. P., and Smith, A. F. M.)*, 1992.

36. L. Leitsalu, T. Haller, T. Esko, M. L. Tammesoo, H. Alavere, H. Snieder, M. Perola, P. C. Ng, R. Mägi, L. Milani, K. Fischer, and A. Metspalu. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol*, 44(4):1137–1147, Aug 2015.

37. Tõnis Tasa, Kristi Krebs, Mart Kals, Reedik Mägi, Volker M. Lauschke, Toomas Haller, Tarmo Puurand, Maido Remm, Tõnu Esko, Andres Metspalu, Jaak Vilo, and Lili Milani. Genetic variation in the estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics*, 27(3):442–454, 2019.

38. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2019.